PERFORMANCE COMPARISON OF RELEVANCE VECTOR MACHINE AND SUPPORT VECTOR MACHINE CLASSSFIERS FOR FUNCTIONAL MRI DATA

A Thesis
Presented to
The Academic Faculty

by

Daniel A. Perez

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Biomedical Engineering

Georgia Institute of Technology
August 2010

PERFORMANCE COMPARISON OF SUPPORT VECTOR MACHINE AND RELEVANCE VECTOR MACHINE CLASSIFIERS FOR FUNCTIONAL MRI DATA

Approved by:

Dr. Xiaoping Hu, Advisor
School of Biomedical Engineering
Georgia Institute of Technology/Emory University

Dr. Shella Keilholz
School of Biomedical Engineering
Georgia Institute of Technology/Emory University

Dr. Charles Isbell
College of Computing
Georgia Institute of Technology

Date Approved: 6/29/2010

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents for their unwavering support and encouragement throughout not only my graduate career, but my whole life. They are the main reason I have become the man I am today. Next, I would like to thank my advisor Dr. Xiaoping Hu for his wisdom and guidance. He has molded me to become a person who knows how to think critically and perform tasks in a timely fashion. These characteristics will be undoubtedly utilized in the next stage of my professional career. Also, I would like to thank my thesis committee for their feedback and input into this work. This work would have no substance without the valuable suggestions from my committee.

Furthermore, I would like to take this time to thank certain BITC members (past and present) who have assisted me in learning the material and in the development of this work. In particular, I would like to thank Dr. Richard Craddock, Dr. Alexandre Franco, and Dr. George James for their indispensable input and conversations. These three served as mentors for me in the lab and aided me in my development as a graduate student. Without them, this thesis certainly would not be possible. In addition, I would like to thank BITC members whose insightful conversations have solidified my understanding of imaging: Steven Harris, Dr. Christopher Glielmi, Alex Poplawski, Changming Wang, Jaemin Shin, Dr. Lei Zhou and Dr. Govind Bhagavatheeshwaran. Finally, I would like to thank my family members and friends who have helped me keep my sanity during these arduous two years of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Functional magnetic resonance imaging (fMRI) is a technique used to measure blood flow associated with neuronal activity. This imaging technique has been used extensively to visualize brain activation. Originally, univariate tests were turned to for the analysis of fMRI data. Recently, multivariate pattern analysis (MVPA) of fMRI data has been growing in popularity due to its sensitivity to networks of brain activation. It is performed in a predictive modeling framework which is natural for implementing brain state prediction and real-time fMRI applications such as brain computer interfaces. Support vector machines (SVM) have been particularly popular for MVPA owing to their high prediction accuracy even with noisy datasets. Recent work has proposed the use of relevance vector machines (RVM) as an alternative to SVM. RVMs are particularly attractive in time sensitive applications such as real-time fMRI since they tend to perform classification faster than SVMs. Despite the use of both methods in fMRI research, little has been done to compare the performance of these two techniques. This study compares RVM to SVM in terms of time and accuracy to determine which is better suited to real-time applications.

CHAPTER 1

INTRODUCTION TO MRI

1.1 A History of MRI

Magnetic resonance imaging (MRI) is a radiology imaging modality used extensively in clinical and research settings to provide an image of the inner structures of the body. This imaging technique has been utilized to provide neurological, musculoskeletal, cardiovascular, and oncological images. In comparison to other common imaging modalities (i.e. X-ray, computed tomography), MRI is a relatively new technology. Its roots can be traced back to the discovery of the nuclear magnetic resonance phenomenon. This pivotal discovery was accomplished independently by Felix Bloch and Edward Purcell in 1946 [1, 2]. In 1952, these founding fathers of magnetic resonance were awarded a Nobel Prize in Physics for their description of the signal produced from the interaction between nuclei and a magnetic field.

In 1971, Raymond Damadian observed that relaxation times between mouse tumors and normal tissue differed in vitro [3]. These results generated great interest in the scientific community since it showed the potential use of the modality for disease detection. In 1973, Paul Lauterbur demonstrated that NMR signals from different locations can be distinguished using magnetic field gradients [4]. Two years later, Richard Ernst and his group proposed magnetic resonance imaging using the Fourier transform [5]. This technique is the foundation of current MRI techniques. Shortly thereafter in 1977, Peter Mansfield developed the echo planar imaging (EPI) technique [6]. This technique still proves to be one of the fastest and most widely used imaging methods. It is still used extensively in neurological applications through functional MRI (fMRI) and diffusion imaging.

In 1991, the Nobel Prize Committee awarded Ernst with a Nobel Prize in Chemistry for his work on the Fourier Transform NMR and MRI. The next year, human fMRI was introduced [7, 8]. This noninvasive imaging technique allowed for visualization of functional activity of the different regions in the brain. In 2003, the Nobel Prize in Medicine was awarded to Mansfield and Lauterbur for their contributions to the field of magnetic resonance imaging. By 2005, it was reported there were approximately 10,000 MRI scanners worldwide. This relatively new field of research will continue expanding and will be repeatedly be used in clinical applications as long as engineers and scientists continue to improve upon the inner workings of the modality.

1.2 Basic MRI Principles

1.2.1 Proton Spin Physics in the Presence of a Magnetic Field

MR imaging is based on the interaction between a spinning proton and an external magnetic field, B_0. Specifically, MRI looks at the nuclei of hydrogen atoms which are known to be found in abundance in the human body residing in water and other molecules. If a proton is placed in a magnetic field, it experiences a torque. The proton attempts to align itself with the external magnetic field but cannot do so. Instead, the proton will precess around the direction of the field (see Figure 1).

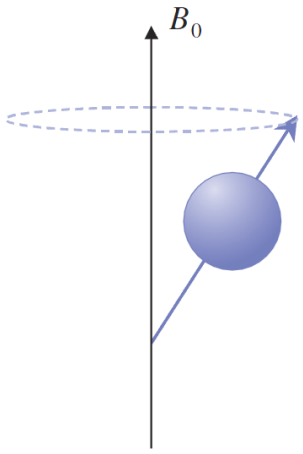

Figure 1: Precession of proton around magnetic field[9]

This precession has a particular frequency which is proportional to the external magnetic field. The precessional frequency can be described using the Larmor equation:

$$\omega_0=\gamma B_0, \quad (1.1)$$

where γ is known as the gyromagnetic ratio and ω_0 is the aforementioned precessional frequency or Larmor frequency. The gyromagnetic ratio is a different value depending on the nuclear species. In water, the hydrogen proton has a gyromagnetic constant of approximately 2.68×10^8 rad/s/Telsa (T). This value converted to Hertz (Hz) yields a gyromagnetic ratio of 42.6 MHz/T. This equation reveals a nucleus will spin in a predictable fashion given a known external magnetic field. For example, for a 2T field, proton spins will precess at a frequency of 85.2 MHz.

Quantum mechanics can also be used to describe the precessional frequency. In the presence of an external magnetic field, the proton will experience a torque making it precess in one of only two orientations or states. One state aligns almost perfectly with the main field, and is known as spin-up or parallel. In contrast, the other state is aligned almost opposite to the main field, and is known as spin-down or anti-parallel (see Figure 2).

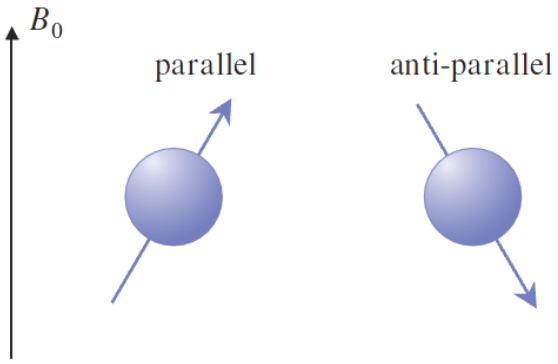

Figure 2: Two possible orientations of protons in an external magnetic field[9]

Both states are stable but differ in energy. The anti-parallel direction requires slightly more energy than the parallel direction. Protons can alternate between the two states by either gaining or losing energy in the form of a photon. The energy of the photon equals

the energy difference between the two orientations. From De Broglie's equation, the relationship between frequency and energy is given by:

$$E=h\omega, \tag{1.2}$$

where h is Plank's constant ($6.626x10^{-34}$J s) and ω is frequency. Furthermore, the energy difference between the two states is known to be as:

$$E=\gamma hB, \tag{1.3}$$

When the two equations are combined, the terms cancel out to yield Larmor's equation. This link between the classical mechanics and quantum mechanics explanation reveals that the precessional frequency of a proton in a magnetic field is the same frequency of the radiation required to change between orientations.

The population of protons favors the lower energy orientation. Therefore, protons tend to spin in the parallel state rather than in the anti-parallel state. The ratio between these two states is dependent on both the main magnetic field strength and ambient temperature. This ratio is termed 'spin excess' and can be calculated using the following formula:

$$\text{Spin excess} \cong N\frac{\hbar\omega_0}{2kT}, \tag{1.4}$$

where N is the total number of spins present in the sample, T is the absolute temperature, ħ is Planck's constant divided by 2π, and k is Boltzmann's constant (1.39×10^{-23} JK^{-1}). For a magnetic field strength of 0.3 T at room temperature, the spin excess is only one in a million spins. This means that for every million protons in the spin-down or anti-parallel direction there are a million and one protons in the spin-up or parallel direction. It is apparent that the spin excess is significantly smaller than the total number of spins. This observation might be misleading and lead to an assumption that no significant signal can be detected. However, the tissues being imaged contain an Avogadro's number of protons which have an average magnetic dipole density or longitudinal equilibrium

magnetization. This net magnetization M_0 aligns directly with the external magnetic field B_0 and is calculated to be on the order of microtesla (μT) (see Figure 3).

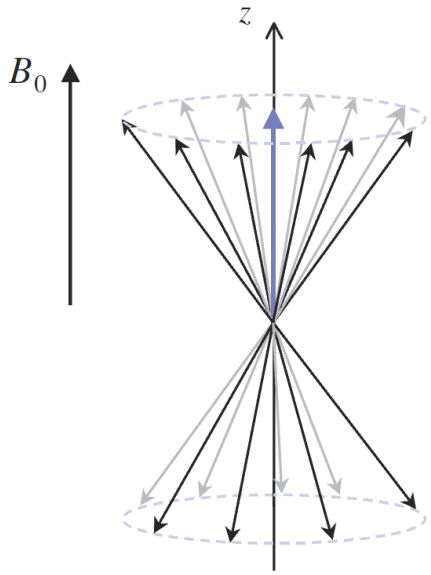

Figure 3. Net magnetization M_0 (in blue) aligning with external magnetic field B_0 [9]

The net magnetization is given by a slight manipulation of the 'spin excess' formula:

$$M_0 = \frac{\rho_0 \gamma^2 \hbar^2}{4kT} B_0, \tag{1.5}$$

where ρ_0 is defined as the number of protons per unit volume or 'spin density'. This equation reveals that the net magnetization generated from a population of spinning protons is extremely small but measurable.

1.2.2 RF Pulse and Flip Angle

As mentioned above, the net magnetization M_0 is relatively small as compared to the external magnetic field B_0. At equilibrium (lying parallel with the external field), the magnetization cannot be detected. Assuming that B_0 lies in the z-direction, the magnetization vector must be tipped away from the external field direction into the transverse plane or x-y plane to get a measurable signal. This tipping into the transverse plane is accomplished using a small duration radiofrequency magnetic field, or rf pulse,

produced by a nearby rf transmit coil. The rf pulse is applied at the Larmor frequency and is therefore said to be in resonance with the proton spins. The degree to which the proton spins are tipped is known as the flip angle and is proportional to the duration of the pulse and the strength of the RF magnetic field B_1. If the proton spins are tipped exactly into the transverse plane, the rf pulse is known to be a 90° pulse. To achieve a 180° pulse, the duration of the pulse could be doubled. However, time is a constraining factor in most MRI applications, therefore, the strength of the RF field is typically manipulated instead to achieve wanted flip angles. After the RF pulse is turned off, the protons spins return to their equilibrium state and release energy in the form of RF waves. This signal is detected by a nearby receiver coil. This is the basic foundation of MRI signal detection.

1.2.3 Relaxation and Bloch Equations

After the application of the RF pulse, the net magnetization is known to return back to its equilibrium state. This transition from excited state back to equilibrium state is known in MRI physics as relaxation. There are two different types of relaxation: spin-lattice relaxation and spin-spin relaxation. The former refers to the exchange of the spins' energy with their surroundings to reach the equilibrium state in alignment with the external magnetic field. As mentioned before, the net magnetization aligns with the external magnetic field during equilibrium. At this point, the z-component of the magnetization vector (M_z) equals the equilibrium net magnetization M_0. M_z is also referred to as the longitudinal magnetization. At equilibrium, the x-component and y-component of the magnetization (M_x and M_y respectively) have no contribution to the net magnetization. After being tipped into the x-y plane, M_z approaches zero. However, relaxation will lead to a gradual increase in the longitudinal magnetization until

realignment with B_0 occurs. This growth of longitudinal magnetization is dependent on the spin-lattice relaxation time or T_1. This value is the time it takes the magnetization to recover 63% of its equilibrium value. Moreover, this T_1 value is highly dependent on the tissue being imaged.

In contrast to spin-lattice relaxation, spin-spin relaxation is related to a reduction in transverse magnetization. This relaxation occurs because proton spins experience local fields which are a combination of the applied field and the fields of their neighbors. As a result, spins will be at different precessing frequencies and will 'fan out' or dephase in time. This dephasing will result in a decay of the transverse magnetization (see Figure 4).

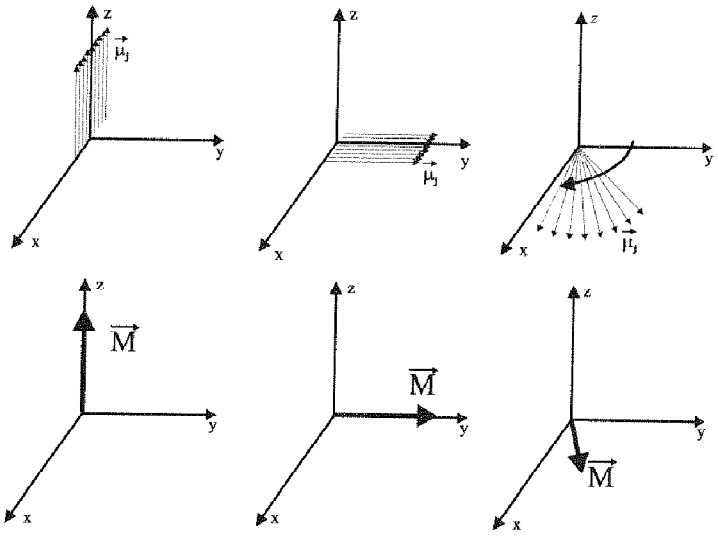

Figure 4. Transverse magnetization. Top row shows dephasing of individual spins after 90 degree tip. Bottom row shows a decrease in the net transverse magnetization because of the dephasing. [10]

The overall reduction of the transverse magnetization refers to a reduction in the x and y-components of the magnetization (M_x and M_y). This rate of decay is dependent on the

'spin-spin' relaxation time or T_2. This value is determined to be the time it takes for the transverse value to drop to 37% of its initial size. In contrast to T_1, T_2 happens very quickly. The transverse magnetization is zero after a few milliseconds while it takes a few seconds before the longitudinal magnetization aligns with the external magnetic field in the z-direction. Both relaxations are shown in Figure 5.

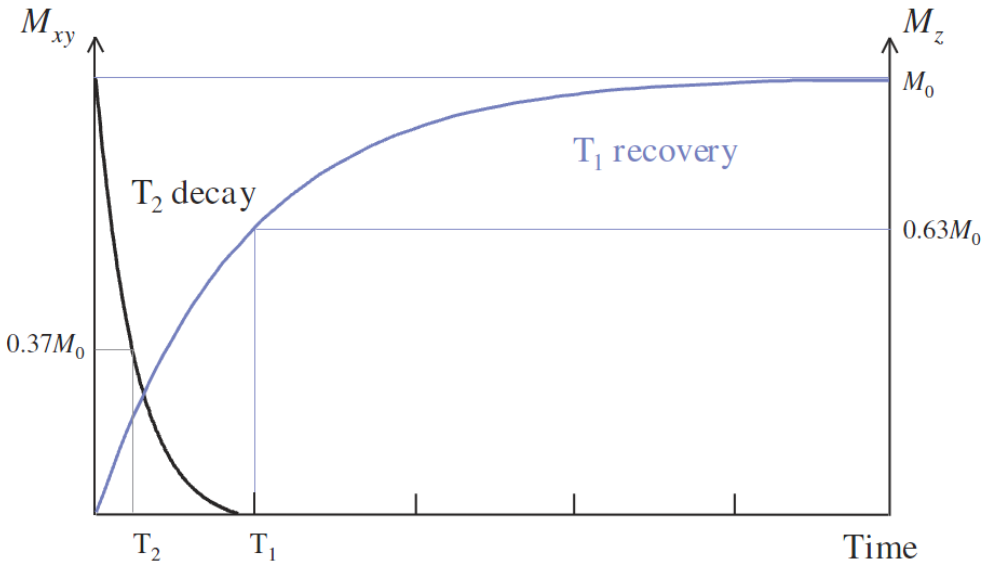

Figure 5. T1 and T2 relaxation [9]

With knowledge of the relaxation rates and external field strength, the dynamics of the magnetization during excitation and relaxation can be described using the Bloch equation:

$$\frac{d\vec{M}}{dt} = \gamma\vec{M} \times \vec{B}_{ext} + \frac{1}{T_1}(M_0 - M_z)\hat{z} - \frac{1}{T_2}\vec{M}_{\perp}, \tag{1.6}$$

where $\vec{M}$ is the magnetization vector, $\vec{B}_{ext}$ is the main magnetic field vector, and $\vec{M}_{\perp}$ is the transverse magnetization vector. For the constant field case or $\vec{B}_{ext}$=$B_0\hat{z}$, the three components of magnetization can be derived from equation 1.6:

$$M_x(t) = e^{-\frac{t}{T_2}}(M_x(0)cos\omega_0 t + M_y(0)sin\omega_0 t) \quad (1.7)$$

$$M_y(t) = e^{-\frac{t}{T_2}}(M_y(0)cos\omega_0 t + M_x(0)sin\omega_0 t) \quad (1.8)$$

$$M_z = M_z(0)e^{-\frac{t}{T_1}} + M_0(1 - e^{-\frac{t}{T_1}}) \quad (1.9)$$

In practice, there are two main factors which contribute to the decay of the transverse magnetization. The first contribution is the molecular interactions which are known as the pure T_2 effect. The second contribution is due to the additional dephasing introduced by variations or inhomogenieties in the B_0 field. These two contributions can be combined to describe the T_2^* relaxation time which accounts for both the external field inhomogeneity (T_2') and pure spin-spin relaxation (T_2). This total relaxation rate can be described by:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'} \quad (1.10)$$

This relaxation rate is considered as the basis of functional magnetic resonance imaging or more specifically the foundation for BOLD (blood oxygen level dependant) contrast. This signal and its implications in functional imaging of the brain will be discussed later.

1.2.4 MR Image Formation

As mentioned before, the precession frequency of spins can be determined at a given magnetic field strength. This relationship is given by Larmor's equation (1.1). Using this equation, it is known that the MR signal at a magnetic field strength of 1.5 T will result in a MR signal with a frequency at 63 MHz. Similarly, an RF pulse applied at this same magnetic field strength will result in MR signals also at 63 MHz. These could

be detected by a coil and receiver tuned to this particular frequency. However, this signal alone cannot produce an image of the patient since there is no way of assigning parts of the signal to where they came from. To accomplish localization of the MR signal, magnetic field gradients are used.

The term gradient refers to a linear variation in the field strength in a particular direction. During moments where a gradient is not present, all the spins experience the same field and have the same frequency. When a gradient is added, the magnetic field produced by the gradient is added to the external field. At the center, the field experienced by the spin is B_0. At this location, the spins will resonate at the Larmor frequency. As one moves away from the center, the spins will either resonate faster or slower since it is experiencing a different field. These differences in precession are detected as higher and lower frequencies in the MR signal. As such, distinctions between MR signals at different positions in space could be made. Three sets of coil systems exist in the MR system each applied in a different orientation: G_x, G_y, and Gz. These gradients allow for three dimensional image acquisitions.

The MR signals obtained using the gradient fields are plotted in a k-space matrix. This k-space holds the raw data before reconstruction. To generate an image, a Fourier transform is applied to the k-space data. The use of the different gradients allow for the filling of the k-space matrix and hence successful image reconstruction. A gradient applied in the x-direction is known as the frequency encode or readout gradient. In contrast, a gradient applied in the y-direction is known as the phase encoding gradient. Furthermore, gradients in the z-direction are termed slice-selection gradients. Signal detection is accomplished during the application of the readout gradient.

Gradients are applied in a controlled fashion through a MR pulse sequence. Pulse sequences differ in timing and presence of RF pulses and gradients. These sequences could dramatically impact the appearance of an image. Pulse sequence choice is dependent on the application at hand. For this particular application, echo planar imaging (EPI) was used. The use of EPI has been method of choice for diffusion, perfusion, and BOLD imaging. This pulse sequence allows for fast imaging but is limited by low resolution and prone to severe artifacts. Despite these limitations, EPI is used extensively for observation of the BOLD signal.

1.3 fMRI Basics: BOLD Contrast and the Hemodynamic response

Functional magnetic resonance imaging (fMRI) is a recent neuroimaging MR technique which measures changes in blood flow related to neural activity. The most common variation of fMRI uses blood oxygen level dependent (BOLD) contrast as its basis for its signal measurement. This valuable fMRI technique was first introduced by Ogawa et al [11]. Oxygen is transported in the blood bound to the protein hemoglobin. When oxygen is bound to hemoglobin (oxyhemoglobin), the molecule becomes diamagnetic. Conversely, when no oxygen is bounded to hemoglobin (deoxyhemoglobin), the molecule becomes paramagnetic. Diamagnetic oxyhemoglobin has no effect on MR signal while presence of paramagnetic deoxyhemoglobin results in local field inhomogeneities and hence a lower MR signal.

During regional increases in neuronal activity, there is a consumption of nearby oxygen and therefore an increase in deoxyhemoglobin. This translates to an initial dip in the fMRI signal. If neuronal activity persists, an increase in blood flow will follow

bringing in new oxygenated blood. The increase of incoming oxygenated blood flow will exceed the rate of oxygen extraction from the blood. This results in higher levels of oxyhemoglobin as opposed to deoxyhemoglobin resulting in a greater fMRI signal. After neuronal activity ceases, blood flow levels return to baseline levels or to an approximately constant ratio of oxyhemoglobin to deoxyhemoglobin. This series of events is collectively known as the hemodynamic response and is used in fMRI as a measure of brain activity. It should be noted that this measurement of brain activity is indirect in nature since it measures local changes in brain vasculature activity. The exact relationship between neuronal activity and its corresponding blood flow is a highly debated topic which has yet to be determined with definite certainty[12].

1.4 fMRI Paradigms and Pre-Processing

Functional MRI studies usually employ one of the two basic experiment designs to study brain activity. The first type of paradigm is known as an event-related design. This setup involves presenting stimuli at randomized times and varying interstimulus intervals. The second experimental paradigm is known as the block design. In this setup, the experiment is divided into alternating periods (blocks) of stimulus presentation and rest or baseline activity. These blocks vary in duration depending on the study and are repeated several times throughout the experiment. Stimulus presentation can be presented visually through the use of a projection device or can be presented in an auditory fashion using special noise cancelling headphones. Complex MR-compatible stimulus presentation devices have been constructed to stimulate other senses such as olfaction and taste [13,14]. Block design paradigms are used extensively during fMRI research due to its simplicity and straightforward statistical analysis. However, the repeating nature of the

blocks can become predictable and potential confounds such as habituation and anticipation could affect the raw results. An example of a block design experiment is presented below in Figure 6.

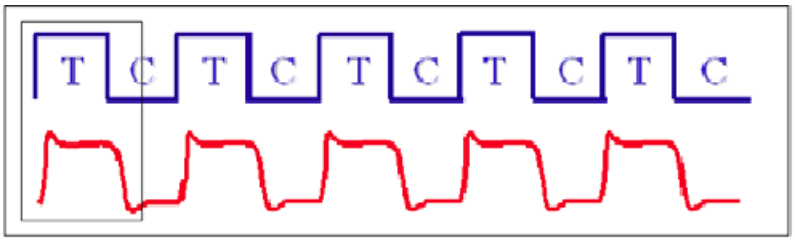

Figure 6. Typical block design paradigm and corresponding hemodynamic response [15]

During a typical fMRI scan, the subject will perform a functional task within the bore of the MRI scanner. The scanner will acquire two-dimensional images which represent slices of a certain thickness of the brain. The slices are combined to form a three dimensional representation of the brain. In practice, various 3-D volumes of the brain are collected over time adding a fourth dimension to the resulting dataset. The time it takes to acquire one brain volume is known as 'time to repetition' or simply 'TR'. This parameter controls the temporal resolution or sampling rate of the experiment. Ideally, one would think that the data should be acquired at a high sampling rate with a low TR to provide the maximum amount of information per unit time. However, lowering the repetition time compromises the spatial resolution of the data. That is, lowering the TR will lessen the number of slices collected resulting in less brain coverage. Fortunately, the temporal characteristics of the hemodynamic response function (HRF) limit the usefulness of using a high temporal resolution. In echo planar imaging, there is no need to collect at TR's lower than 1 second[16].

The spatial resolution of the dataset is dependent on the voxel size of the image. Voxels are volume elements which are similar to pixels but represent a quantity of 3-D data. A voxel represents the smallest amount of brain tissue imaged. The best spatial resolution can be accomplished by using the smallest voxel size possible. Increasing spatial resolution or lowering voxel size could have a detrimental effect on the signal to noise ratio and temporal resolution. Spatial resolution is also affected by a commonly used neuroimaging postprocessing step: spatial smoothing. This procedure facilitates in mapping techniques used in group comparison studies.

After brain image acquisition, several pre-processing steps have to be performed before an accurate analysis of the data could be accomplished. Some general preprocessing steps include slice timing correction, motion correction, spatial normalization of data to a brain template, spatial/temporal filtering, and resampling of data. These preprocessing steps can be done using a variety of software packages which are readily available online. The optimal choice of preprocessing steps and order of the steps has not been defined at this moment and varies among experimenters.

After the preprocessing of the data, the HRF has to be modeled to represent the fMRI signal change in an activated region. The general procedure is to the model the generic shape of the BOLD response using most likely gamma or passion functions [15]. One of the easiest methods to obtain results from a two state fMRI experiment is to use the subtraction technique. All the images during the task blocks are averaged and are subtracted by the average of all the rest images. Although easy to implement, this technique has the disadvantage of yielding observable artifacts due to head motion. A variation of this technique uses a Student's t-test to reduce the amount of artifact seen.

This method factors in the standard deviations of the two states. The t-score is calculated on a pixel-by-pixel basis and compared to a statistical threshold. Any voxel with a t-value greater than this threshold is denoted as active. Various other univariate techniques such as general linear modeling, cross-correlation, and Fourier have all been applied to fMRI for analysis[9, 17]. Similarly, multivariate techniques such as principal component analysis, connectivity analysis, and pattern recognition algorithms have all been used to observe fMRI results [18, 19]. This last technique encompasses using machine learning algorithms like support vector machines and relevance vector machines. These two algorithms serve as the foundation of this thesis and will be described in the next chapter.

1.5 Real-Time fMRI

By simplifying some of the preprocessing steps seen in a conventional fMRI analysis and using faster imaging sequences, real-time functional magnetic resonance (rtfMRI) imaging could be accomplished. Real-time fMRI, introduced by Cox et al in 1995, makes it possible to watch one's brain activation 'live'[20]. The benefits of rtfMRI are primarily seen in the clinical setting. During a clinical scan, an immediate feedback of the image quality, results, and compliance is desirable and is offered by using rtfMRI [21]. Furthermore, biofeedback studies are possible with this modality. Subjects could receive information about their head position or brain activation in 'real time'[22]. Lastly, this technique has been utilized to drive brain-computer interfaces[23].

Despite its great promise in a variety of neuroscientific applications, rtfMRI has a limitation which stems from the nature of the fMRI signal itself. As mentioned before, fMRI measures blood flow rather than neuronal activity. Thus, this technique is indirect

and noisy and is incapable of measuring activity deriving from single neuron. Moreover, the time scale of neuronal activity cannot be picked up by the fMRI technique. As mentioned before, blood flow to an activated region is delayed by a few seconds and with a high spatial spread which encompasses more than a million individual cells. One could think of situations in which this would be undesirable. For example, this limitation could affect the temporal performance of a brain machine interface since the system would detect brain activation via blood flow a few seconds after regional neuronal activity. Some researchers argue that even if it was possible to measure billions of neurons in the brain simultaneously and at a high temporal resolution, this information would be redundant. As such, spatial and temporal averaging would probably be done to the data. This spatial and temporal averaging is inherent to fMRI[24]. Despite the apparent temporal and spatial limitations that exist with this technique, rtfMRI is still a powerful tool used to view mental states in real-time.

CHAPTER 2

MACHINE LEARNING

2.1 Machine Learning Classifiers for fMRI

Recently, researchers have turned to the use of multivariate pattern recognition algorithms in analyzing fMRI data and for real-time applications. Many studies have shown that machine learning classifiers can be used to extract new information from neuroimaging data. A classifier can be defined as a supervised machine learning procedure which takes the values of various features in a set of independent variables and predicts the class that the independent variables belong to.

When applied to fMRI analysis, the features are known to be the voxels while the class is the brain state which varies depending on the stimulus presented. Each brain volume or example, $\mathbf{x}=[x_1 \ldots x_v]$, can be denoted by a series of features (voxels) and a corresponding class label y. A classifier will then be 'trained' on a subset of the data known as the 'training dataset.' The training will prime the classifier parameters allowing it to determine what features used contain information about the classes. The learned classifier is applied to a different subset of the data known as the 'testing dataset.' Theoretically, the classifier should be capable of predicting the classes of datasets it has never seen before [25]. The most common metric for determining the performance of a classifier is accuracy. Accuracy can be determined by simply comparing the predicted labels to the actual labels of the test dataset. If the predicted labels match that of the true labels, the classifier predicted correctly. The number of correct predictions can then be

divided by the number of examples in the test dataset yielding the accuracy of the classifier (Figure 7).

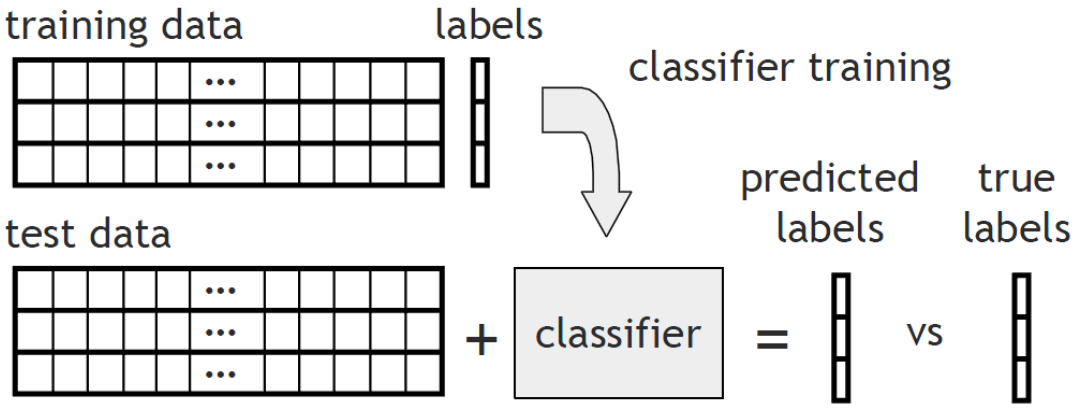

Figure 7.Training and testing of classifier [25]

One procedure which is commonly done before the application of a classification algorithm is the reduction of the number of features. This is termed feature selection. The goal of feature selection is to get rid of superfluous information and to let the classifier concentrate on the features which contain class information. One example of feature selection would be training on voxels within a particular region of interest (ROI). Feature selection is often paired with cross validation. Briefly alluded to earlier, cross validation is a statistical technique which partitions a dataset into subsets. It performs an analysis on one subset (training set) and validates the analysis on the other subset (testing set). Multiple rounds of cross validation are performed using different partitions to reduce variability. The results from each validation of partition are averaged together. One common cross validation scheme is leave-one-out cross-validation (LOOCV). This variant of cross-validation calls for the exclusion of one observation from the whole dataset. The training is done on the remaining data and the classifier can then be tested on

the removed observation. This procedure is repeated until each observation has been excluded and tested.

Upon training, there may be instances in which the classifier can classify the examples seen in the training dataset well but may perform poorly with the testing dataset. This may be a result from the phenomenon known as overfitting [25]. This problem could be addressed by the use of a simple function such as a linear classifier. For fMRI brain state classification, a successful classifier should be able to distinguish between spatial patterns measured under different mental states. For example, let's take a two voxel image of a brain. The fMRI signal within these voxels can be recorded throughout a scan with a presentation of two different visual stimuli. Since this is a two-dimensional case, each measurement of brain activity can be plotted on a Cartesian plane having the x-axis represent voxel 1 values and the y-axis represent voxel 2 values. A classifier could be trained using this data and be used to draw a line between the two classes representing the visual stimulus presented (Figure 8). The classifier would be considered effective if it could accurately predict the corresponding visual state for a new set of voxel values. These classifiers have been proven to work well with multidimensional datasets (i.e. fMRI). Two particular classifiers of interest are known as support vector machines and relevance vector machines. The theory behind both of these pattern recognition algorithms is presented below.

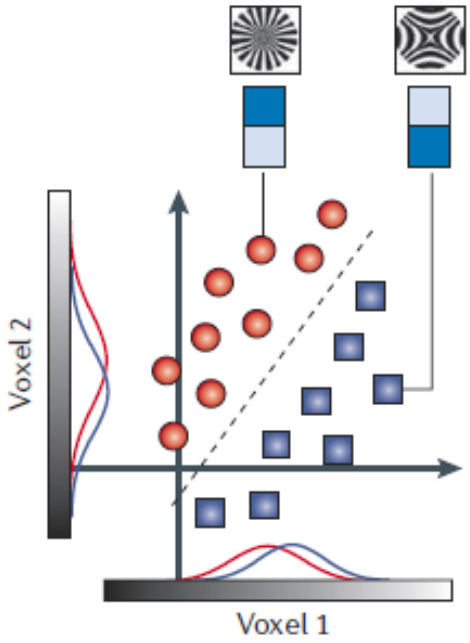

Figure 8. Linear classifier distinguishing mental states using two voxels [26]

2.2 Support Vector Machines

Support Vector Machines (SVM) are a set of supervised pattern recognition algorithms used in a variety of applications for two-group classification problems. This machine learning method was introduced by Vapnik and Cortes in 1995[27]. A brief description is provided in this section.

An SVM model constructs a hyperplane or set of hyperplanes which attempts to separate data into two categories. The best separation between the two categories is achieved by the largest distance to the nearest training data points. Suppose there are l observations. Each observation consists of a vector $\mathbf{x}_i \in R^n$, where i=1,..,l and the associated label, y_i. The label could either be 1 or -1. For a typical fMRI analysis, $\boldsymbol{x}_i$ could be a vector of voxel values and y_i could represent a brain state. SVM attempts to

construct a maximum margin hyperplane that divides the two classes. The points x which lie on the hyperplane satisfy:

$$\mathbf{x} \cdot \mathbf{w} + \mathbf{b} = \mathbf{0} ,$$

(2.1)

where **w** is a vector normal to the hyperplane , b/‖**w**‖ is a parameter that determines the offset of the hyperplane from the origin along the vector **w**, and ‖**w**‖ is the Euclidian norm or magnitude of **w**. The parameters **w** (weight) and b (bias) are chosen to maximize the margin while still maintaining separation of the data. If the data is linearly separable, the margins can be chosen so that no data points lie between them. The two margins can be described by these following constraints:

$$\mathbf{x_i} \cdot \mathbf{w} + \mathbf{b} \geq \mathbf{1} \text{ for } y_i = 1 \quad (2.2)$$

$$\mathbf{x_i} \cdot \mathbf{w} + \mathbf{b} \leq -\mathbf{1} \text{for } y_i = -1 \quad (2.3)$$

The distance between the two margins is simply 2/‖w‖. To maximize this distance, w should be minimized. An example two-dimensional solution is shown in Figure 9. The data points which lie on the margins are known as support vectors. Their removal would lead to a different solution.

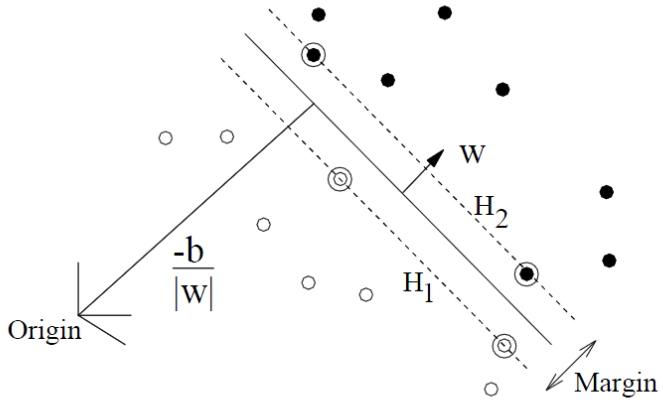

Figure 9. Hyperplane for sample separable data. Support vectors which lie on margins are circled [28].

To compute **w** and b for a maximum margin hyperplane, the following optimization problem has to be solved:

$$\text{Minimize } \frac{1}{2}\|w\|^2 \tag{2.4}$$

$$\text{Subject to: } y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1, \text{ for } i=1,..,n$$

These set of formulas above describe a quadratic optimization problem that could be solved using standard tools from convex optimization.

The example presented above is known as the hard margin SVM and is applicable to linearly separable data. In practice, data are often not linearly separable. Even if they are separable, a much greater margin can be achieved if some misclassification occurs. To allow for some error, a slack variable is introduced:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i, \text{ for } i=1,..,n, \tag{2.5}$$

where $\xi_i \geq 0$ are slack variables that allow some training misclassification. The excess use of slack variables is minimized by a parameter C during optimization:

$$\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi i \tag{2.6}$$

$$\text{Subject to: } y_i(\, x_i \cdot w + b) \geq 1 - \xi i,$$

$$\xi_i \geq 0\,.$$

This variation of SVM is known as soft-margin. This variant of SVM for nonseparable data is affected by the free parameter C which controls the tradeoff between zero training error and increased flexibility brought by the slack variables. A small value for the parameter C will result in many training errors while a large value for the free parameter will result in a behavior similar to that of a hard margin SVM [29].

Lagrange multipliers can be used to simply the optimization problem presented in 2.5. This results in a dual representation of the maximum margin problem:

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j < x_i, x_j > \tag{2.7}$$

subject to:

$$0 \leq \alpha_i \leq C, \tag{2.8}$$

$$\sum_i \alpha_i y_i = 0. \tag{2.9}$$

Quadratic programming can then be subsequently applied to the above problem to train the model[28]. A non-linear classifier could be created using the kernel trick. Equation 2.6 is reformulated and the dot product is replaced by a kernel function $k(x_i,x_j)$.This technique allows for the algorithm to fit the maximum margin hyperplane in a transformed higher dimension feature space. Kernel functions can range from being linear, quadratic, or a Gaussian radial basis function [30].

In this study, SVM was implemented using the SVM routines of MATLAB's bioinformatics toolbox (Mathworks, Natick RI). This toolbox has been used for a wide range of applications including genetics [31], speech recognition [32], and medical

applications. More importantly, this toolbox has been shown to work with fMRI data as well [33].

2.3 Relevance Vector Machines

The SVM classifier is known to be quite successful in a multitude of applications. However, there are a number of significant and practical disadvantages that exist with the use of this method:

- The number of support vectors grows linearly with the size of the training set.
- Predictions are not probabilistic and hence, SVM is not suitable for classification tasks in which posterior probabilities of class membership are necessary.
- It requires a cross-validation procedure to estimate the parameter C which is a waste of data and computation.
- The kernel function must satisfy Mercer's condition; therefore, it must be a continuous symmetric kernel of a positive integral operator.

All of these problems are rectified by using relevance vector machines (RVM) which was originally developed by Tipping [34]. This algorithm is similar in function to SVM but adopts a Bayesian approach to learning. The use of RVM's requires dramatically fewer kernel functions which could lead to significant reduction in the computational complexity of the decision function, therefore, making RVM more suitable for real-time applications. In contrast to SVM, the non-zero weights in the RVM are not associated with examples near the boundary, but rather represent prototypical examples of the

classes. A brief description of the theory behind the relevance vector machine is presented below. For a more detailed explanation, one should refer to [34] and [35].

The RVM algorithm is based on a linear model where predictions are made based on a function of the form:

$$y(x_n) = \sum_{m=1}^{M} w_m \phi_m(x_n), \tag{2.10}$$

where $w_m = (w_1,\ldots,w_m)^T$ are the model weights and $\boldsymbol{\phi_m}$ represents the basis functions. The elements of the RVM basis functions can be given by kernels:

$$\phi_m(x_n) = k(x_n, x_m). \tag{2.11}$$

Through training of the RVM, the optimal weights are determined. Model complexity, over-fitting and computational expenses are controlled by setting the weights to zero thereby inducing sparsity.

In RVM regression, assume one is given a training set consisting of paired input data and real valued input targets: x_n and t_n respectively, for n=1,...,N. For the training dataset, the target vector can be expressed as:

$$t_n = y(x_n) + \epsilon, \tag{2.11}$$

where $\boldsymbol{\epsilon}$ is the noise vector with an assumed zero mean Gaussian distribution and variance of σ^2. Thus, $p(t_n|x) = \mathbf{N}(t_n|y(x), \sigma^2)$ where the notation denotes a Gaussian distribution for t_n with mean $y(x_n)$ and variance σ^2. Due to the independence of t_n, the likelihood of the dataset can be written as:

$$p(t|w,\sigma^2) = (2\pi\sigma^2)^{-N/2}\exp\{-\frac{1}{2\pi\sigma^2}\left||t-\Phi w|\right|^2\}, \quad (2.12)$$

where Φ is the N x M design matrix with $\Phi = [\phi(x_1), \phi(x_2), \ldots, \phi(x_N)]^T$.

Since the number of parameters in the model matches that of training examples, the maximum likelihood estimation of w and σ^2 would lead to overfitting. In RVM, this problem is circumvented by the use of an explicit prior distribution over the parameters. A zero-mean Gaussian prior distribution is introduced over the weight parameters:

$$p(w|\alpha) = \prod_{i=0}^{N}\mathbf{N}(w_m|0,\alpha_m^{-1}), \quad (2.13)$$

where the independent hyperparameters, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$ control the strength of the priors over the corresponding weights. This key feature of the model is responsible for the sparsity seen in RVM. Using Bayes' Rule, the posterior over the weights can then be given by:

$$p(w|t,\alpha,\sigma^2) = (2\pi)^{-(N+1)/2}|\sum|^{-1/2}\exp\{-\frac{1}{2}(w-\mu)^T\sum^{-1}(w-\mu)\}, \quad (2.14)$$

with

$$\sum = (\Phi^T B\Phi + A)^{-1}, \quad (2.15)$$

$$\mu = \sum\Phi^T Bt, \quad (2.16)$$

where $A = diag(\alpha_1, \ldots, \alpha_M) and\ B = \sigma^{-2}I_N$. It should be noted that σ^2 should be treated as a parameter and estimated from the data

Finally, the weights can be integrated out and the marginal likelihood for the hyperparameters can be obtained. Optimization of the marginal likelihood with respect to α and σ^2 is performed using an interpolation procedure introduced by MacKay [36]. Predictions can then be made using these maximizing values.

In RVM classification, suppose one is given a dataset consisting of an input and its corresponding class label: $\{x_n, t_n\}$ for $n=1,\ldots,N$ where $t_n \in [0,1]$. For classification purposes, the model is a linear combination of basis functions transformed by a logistic sigmoid function:

$$y(x, w) = \sigma\{w^T \phi(x)\}, \tag{2.17}$$

$$= \frac{1}{1+\exp\{w^T \phi(x)\}}. \tag{2.18}$$

Adopting the Bernoulli distribution for P(t|x), the likelihood can be written as:

$$P(t|x) = \prod_{n=1}^{N} \sigma\{y(x_n; w)\}^{t_n}(1 - \sigma\{y(x_n; w)\}^{1-t_n}. \tag{2.19}$$

In contrast to the previous regression explanation, the weights cannot be integrated out analytically. As a result, there are no closed form expressions for neither the weight posterior p(w|t,α) or marginal likelihood P(t|α). Therefore, a Laplacian approximation procedure is utilized.

Since $p(w|t,\alpha) \propto P(t|w)p(w|\alpha)$, the maximization can be done by taking natural logarithm of both sides:

$$\ln p(w|t, \alpha) = \sum_{n=1}^{N}(t_n \ln y_n + (1 - t_n) \ln(1 - y_n)) - \frac{1}{2} w^T A w. \tag{2.20}$$

This is a logistic log-likelihood function and requires an iterative maximization. To accomplish this, an iteratively-reweighed least-squares (IRLS) algorithm is used [37]. The quantity in 2.17 is twice differentiated with respect to w to yield the Hessian matrix :

$$H = \nabla_w \nabla_w \ln p(w|t, \alpha) = -(\Phi^T B \Phi + A), \tag{2.21}$$

where B=diag ($\beta_1, \ldots, \beta_N$) with $\beta_n = \sigma\{y(x_n)\}^{tn}[1- \sigma\{y(x_n)\}]^{1-tn}$. This result is negated then inverted to give the covariance $\sum$ for a Gaussian approximation centered to the posterior weight centered at the most probable weight w_{MP}. Using the covariance statistic and w_{MP}, the hyperparameters α_i can be updated:

$$\alpha_i^{new} = \frac{1-\alpha_i \sum_{ii}}{w_{MP}^2}, \tag{2.22}$$

where $\sum_{ii}$ is the ith diagonal element of the covariance. During optimization, many α_i will have large values. This corresponds with the pruning out of the corresponding model weights resulting in sparsity. As mentioned above, those non-zero examples are termed relevance vectors and seem to represent typical instances of classes rather than the borderline possibly misclassified instances support vectors found in SVM.

In this study, RVM was implemented using Michael Tipping's Sparse Bayesian learning toolbox for MATLAB (www.miketipping.com). Relevance vector machines have been used in a variety of applications such as cancer detection [38], bioinformatics [39], and even environmental studies [40]. Recently, the use of this algorithm has been extended to neuroimaging (i.e. rtfMRI studies) [41].

CHAPTER 3

METHODOLOGY

3.1 BOLD Simulation

Before the pattern recognition algorithms could be tested on fMRI data acquired from human subjects, a simple simulation was conducted to test the attributes of both SVM and RVM on high dimensional data in a controlled environment. This simulation was conducted using MATLAB. A simple ten voxel model of a brain was constructed. Each voxel had its own corresponding BOLD signal which only differed in magnitude but not phase. This BOLD signal was constructed by convolving the stimulus paradigm with a model of the hemodynamic response function.

The stimulus paradigm was similar in structure to the one used during the human fMRI analysis. The total length of the stimulus waveform was 6 minutes and it consisted of 12 epochs of 30 second duration. Each epoch consisted of two states corresponding to rest and activation, or 0 and 1 respectively. The TR used was 2 seconds which is similar to the one used for the later studies. As a result, each state in an epoch contained 15 scans. A model hemodynamic response function was created using AFNI using a TR or sampling rate of 2 seconds [42]. The convolution of the two above waveforms results in an ideal hemodynamic response to a block design stimulus setup of rest and activation (Figure 10).

To test the capabilities of the algorithms, increasing levels of noise were added to the voxels' BOLD signals using a random number generator. The amount of noise added

was kept consistent among all the voxels. Leave one out cross validation (LOOCV) was performed by extracting one epoch and designating it as the testing data. Training was subsequently done on the rest of the data using the pattern recognition algorithms. This was done until all the epochs were tested. The classifier output was detrended to eliminate classifier drift. It should be noted that for both models a linear kernel was implemented. Furthermore, for SVM, the parameter C was set to 1 for all cases. Typically, the performance of the support vector machine only degrades at very small C values [43]. For RVM, the initial α was kept at its' default value. The manual accompanying the software revealed that the algorithm was not sensitive to this choice as long as no extreme values were used. The kernel width ξ was set at 5.

For each validation, the accuracy, number of vectors per number of training scans, training time, and prediction time were noted. Classification accuracy is quantified simply by dividing the number of correctly classified scans by the total number of scans. This result is then multiplied by 100.

Since a random number generator was used as a source of noise, the above procedure was done 10 times to insure the results seen were an indicator of the algorithms' performance and not a byproduct from the noise. The metrics recorded were averaged across the 10 runs. The analysis was performed using three different techniques: keeping all the training and testing scans, discarding the transitional scans in the training data but not the testing data, and discarding the transitional scans in both datasets. Transitional scans refer to the few scans between the two states in which the BOLD signal is either rising or falling. Removal of these transitions for both datasets is common practice for offline studies [44]. In this study, three transitional images are removed. The

BOLD response is known to hit a plateau after 6-9 seconds after a stimulus onset [45]. Thus, using a TR of 2 seconds, the exclusion of 3 images insures that the datasets contain only maximal activations.

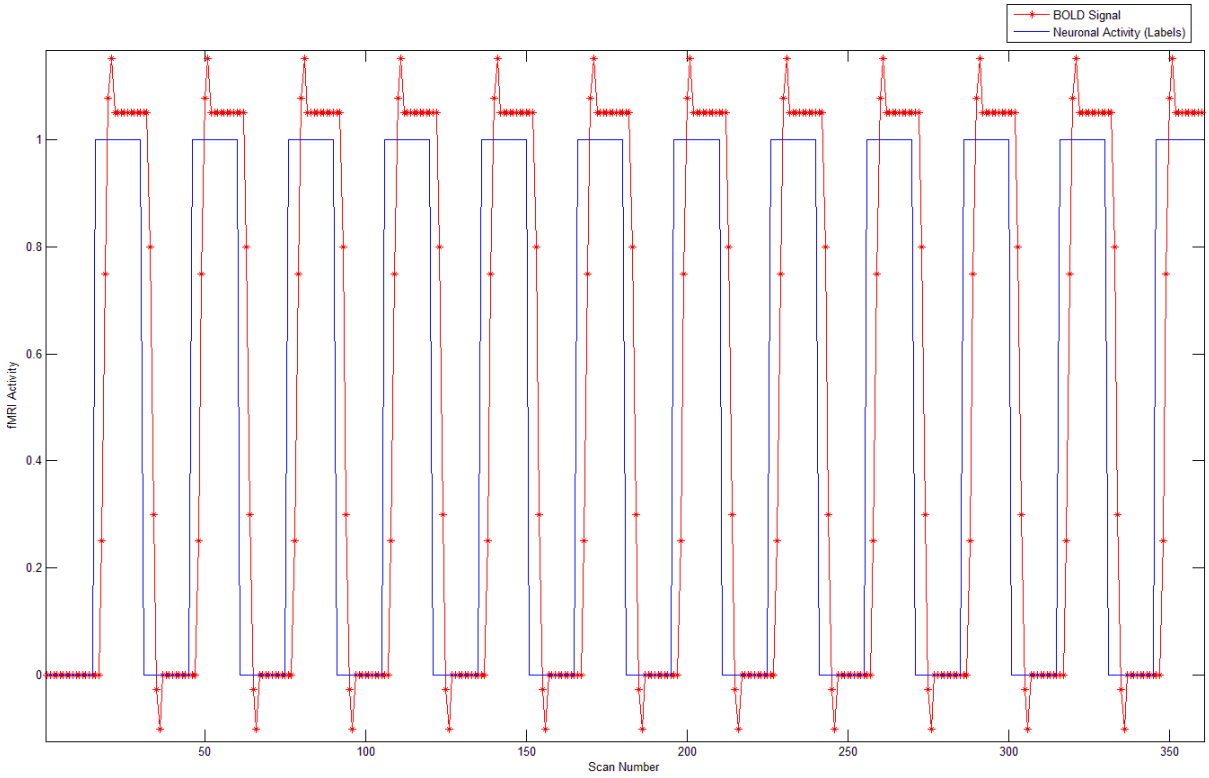

Figure 10. Simple block design paradigm (blue) and its corresponding ideal BOLD signal (red)

3.2 Real fMRI Data

3.2.1 Subjects

A total of eight healthy participants (5 males, 3 females) were examined in this study in accordance with IRB policy. These subjects were all right handed. All of the subjects gave informed consent and were compensated for their time.

3.2.2 Equipment and Software

All scans were performed with the use of a Siemens Magnetom Trio 3T MRI Scanner using a 12 channel head matrix coil. Participants were given earplugs to protect them from the scanner's inherently high noise levels. MR-compatible air conducting headphones were given to the subjects so that the operator could communicate with the subject. Moreover, subjects were given a button controller for use during the motor paradigm. The stimulus was presented using Presentation® (www.neurobs.com). Data-preprocessing was accomplished using AFNI and in house MATLAB scripts. As mentioned before, both SVM and RVM classification were run in MATLAB using the bioinformatics toolbox and Tipping's original software implementation respectively.

3.2.3 Scanner Parameters

Before any fMRI data could be acquired, a T1-weighted anatomical image was acquired using an MPRAGE sequence. Functional MRI results were later overlaid onto this acquired high resolution T1 image for visualization purposes. The imaging parameters for the anatomical scan were as follows: field of view (FOV) =256 mm, repetition time (TR)/ echo time (TE) / flip angle (FA) = 2.6 s/3.02 ms/8^{o} and resolution= 1 x 1 x 1 mm^{3}. For the BOLD contrast data acquisition, an echo planar imaging sequence was used. The imaging parameters were as follows: FOV=192mm, TR/TE/FA = 2s/28ms/90^{o}, resolution=3x3 mm, and 34 3 mm-axial slices. Four 6- minute runs were done using these parameters to measure the BOLD signal resulting from periods of rest and stimulus presentation.

3.2.4 Stimuli

Visual stimulation was implemeted for two runs while monitoring of motor activity was done for the remaining other runs. The visual stimulus consisted of alternating blocks of visual stimulus (an 8 Hz flashing checkerboard) and rest (static fixation cross). Each state lasted 30 seconds (15 images with the given TR of 2 seconds). A total run yielded 180 images containing equal amounts of each state.

The motor paradigm had interleaving periods of rest (static fixation cross) and motor activity (right or left arrow presentation). Each run only stimulated one motor direction, either right or left. Before the scan, the subjects were instructed to press the button with their right thumb if an arrow pointing to the right was shown. The left thumb was used for left arrow presentation. The duration of the run and the number of images of each state was the same as in the visual case. A subset of the basic paradigm is presented in Figure 11.

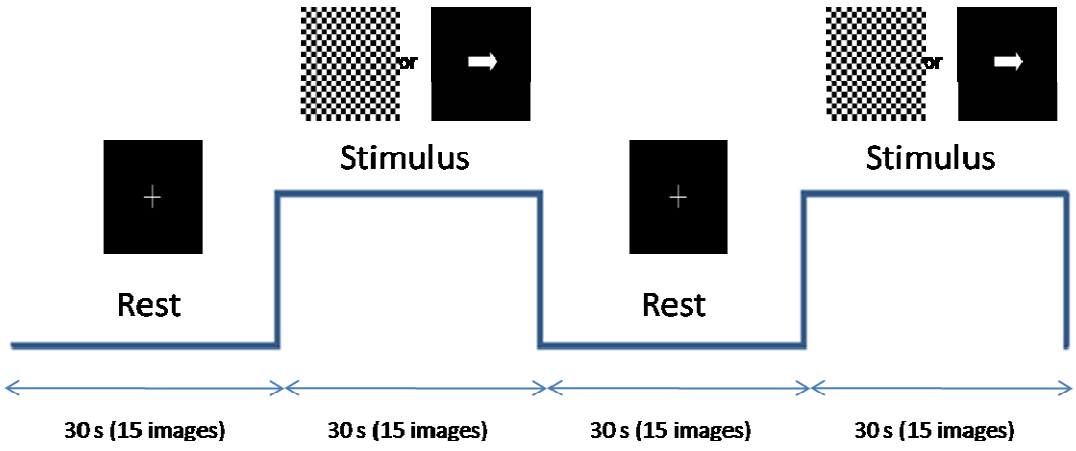

Figure 11. Paradigm used for visual and motor stimulation

3.2.5 Data Preprocessing: Single Subject

Both the anatomic and fMRI images were evaluated for common imaging artifacts such as ghosting, banding, and other imaging errors. No images were removed in this study due to artifacts. Images were corrected for slice timing and motion. A mask encompassing the whole brain was constructed and applied to the datasets to reduce the number of features only to the voxels within the brain. The datasets were next detrended and standardized using z-score. The two visual runs were concatenated together to form one large dataset comprising of 360 images.

3.2.6 Data Preprocessing: Multi-subject

In addition to the steps in 3.2.5, new datasets were formed for muti-subject analysis by transforming both the anatomical and fMRI images into Talairach space [46]. Next, the fMRI images were spatially smoothed using a 6-mm FWMH Gaussian kernel. This smoothing is necessary for the multi-subject case to lessen the normally high intersubject anatomical variability[47].

3.2.7 Feature Selection

Filter feature selection (FS) was implemented by first smoothing the dataset using an 6-mm FWMH Gaussian kernel and applying a general linear model (GLM) using a hemodynamic model of the stimulus task as the regressor of interest. It should be noted that for the multi-subject case, smoothing is not necessary. GLM results were thresholded ($q=0.005$) and voxels passing this threshold were retained for further analysis by SVM and RVM.

3.2.8 Classification: Single Subject

The classification process implemented in this section was very similar to the one done during the simulation phase. The same model parameters were used for this classification. To summarize, LOOCV was performed by extracting one epoch containing 15 images of both rest and stimulation while the training the SVM or RVM model on the remainder of the data. This process was done to test every extracted epoch. Classifier drift had to be accounted for through detrending. For each validation, performance metrics were recorded (i.e. prediction accuracy, training time, etc). For the visual paradigm, the resulting concatenated dataset contained 360 images. For each motor paradigm, the number of images was 180. Therefore, in performing LOOCV in the visual scenario, 12 different combinations of training/testing data were modeled per subject. In contrast, 6 different combinations of training/testing data were modeled per subject for each motor run.

This classification process was done for two situations using whole brain data and using voxels passing the FS filter: training/testing with no removal of transitional images and training/testing with removal of transitional images in the training subset. The additional scenario found in the simulation involving the removal of transitional images in both subsets of data was left out because this is typically done only in offline studies. The potential use of these algorithms is mainly for real-time applications in which the brain state during the testing phase is unknown. Therefore, the presence of transitions would also be unknown. Removal of these volumes would not mirror the environment seen during rtfMRI.

3.2.9 Classification: Multi-subject

Multi-subject classification is accomplished using the smoothed dataset transformed into Talairach space. The visual images for all the subjects were concatenated together. The same process was applied to both the motor right and left paradigms. LOOCV was implemented by removing one subject's data and training on the rest of the subjects. Testing was done for each subject using whole brain data and voxels passing the FS threshold. Comparisons were also drawn observing the effect of the removal of transitional images from the training dataset. For this analysis, the same model parameters used in sections 3.1 and 3.2.8 were applied.

3.2.10 Visualization of Weights

For each case, the weights were visualized by mapping the weight vector back to the voxel space. This was done to show the importance of the voxels to the prediction task. The weight value of each voxel in the weight vector indicates the importance of the voxel in discriminating between the two classes [48]. For SVM, the model's alpha after training was multiplied by the support vectors to yield a weight vector. For RVM, there was no precedence to visualize the weights. However, a similar approach was used and the model's training weights were multiplied by the relevance vectors. After this matrix multiplication, the results were z-scored and the absolute value was taken. The features were ranked and the top ten percent were mapped back onto the T1 image.

CHAPTER 4

RESULTS

4.1 Simulation results

In this section, performance metrics were recorded for SVM and RVM using simulated high dimensional fMRI data with varying amounts of noise. The effects of noise addition was observed for three cases: training/testing without removal of transitional images, training/testing with removal of transitional images from the training subset, and training/testing with removal of transitional images from both subsets of data. In Figures 12-15 the effects of adding noise to the signal can be seen for different performances measures. A legend is presented in Figure 16 representing each scenario.

Figure 12 reveals the effects of a noisy dataset on prediction accuracy. Prediction accuracies dramatically fell with increasing noise levels for both support vector machine and relevance vector machine classification. The removal of transitional images from both training and testing yielded the lowest misclassification rates ranging from approximately 95 percent at low noise levels to about 80 percent at high noise levels. The other two training/testing scenarios yielded slightly lower accuracies. All scenarios had classification accuracies that were better than chance (50%). The lowest accuracies were seen at high level noise levels when classifications using whole datasets were performed.

The effect of noise on the percentage of vectors used for training during classification was observed in Figure 13. At low noise levels, both algorithms use a relatively small amount of vectors for training. The only exception is found in SVM

classification using whole datasets where the percentage of vectors used for training fluctuates close to 100 percent regardless of noise level. As noise is added, the number of vectors increases dramatically for each SVM case to 67 percent for classification without training transitional images and 51 percent for classification without transitional images in both subsets. The increase of noise does not seem to affect amount of vectors used for RVM classification. At high noise levels, this percentage stays between 1.1-1.4 % of the training data.

Training times and prediction times were inspected in Figures 14 and 15 respectively. The effects of noise are different for each algorithm. Low noise levels result in high training times for RVM classification in relation to SVM classification. Increasing noise decreases the training time for RVM dramatically from low to high noise while slightly lowering training times for SVM. Furthermore, the addition of noise increases the decision time for SVM. The decision time for RVM classification is much lower than that of SVM for all cases and does not seem to change with added noise. These findings are very similar to the waveforms seen in Figure 13. This suggests correspondence between sparsity and prediction times.

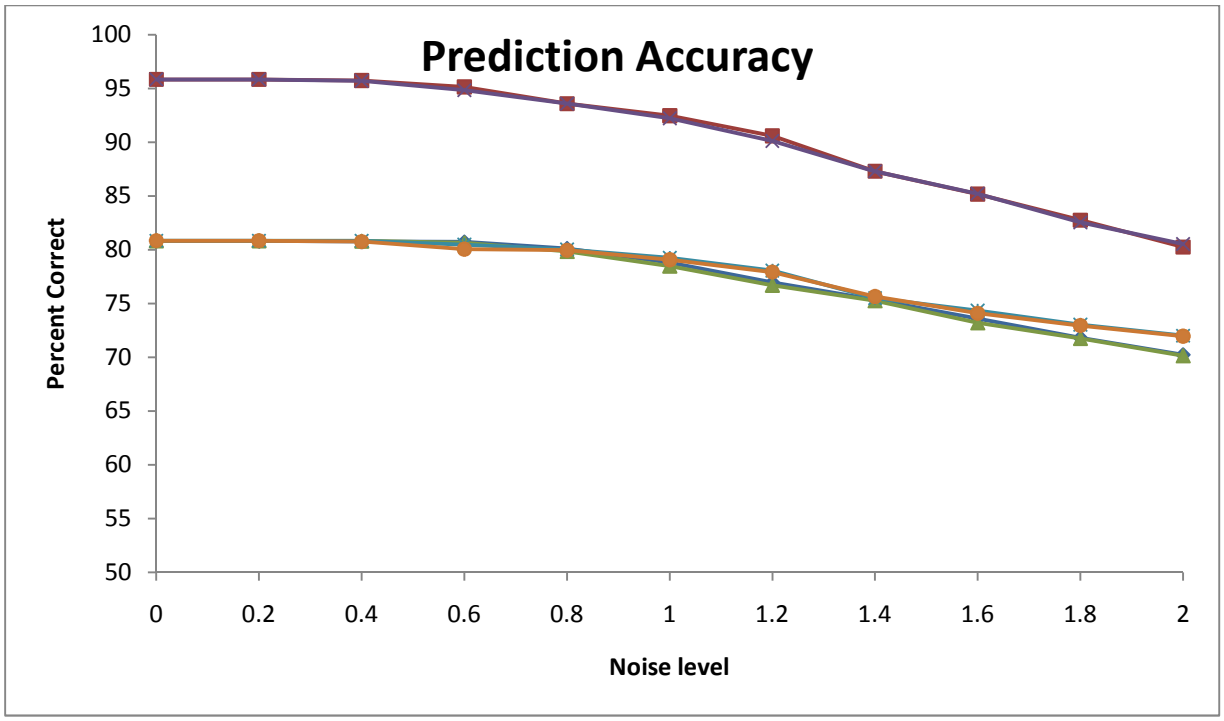

Figure 12. Prediction accuracies for simulated fMRI data

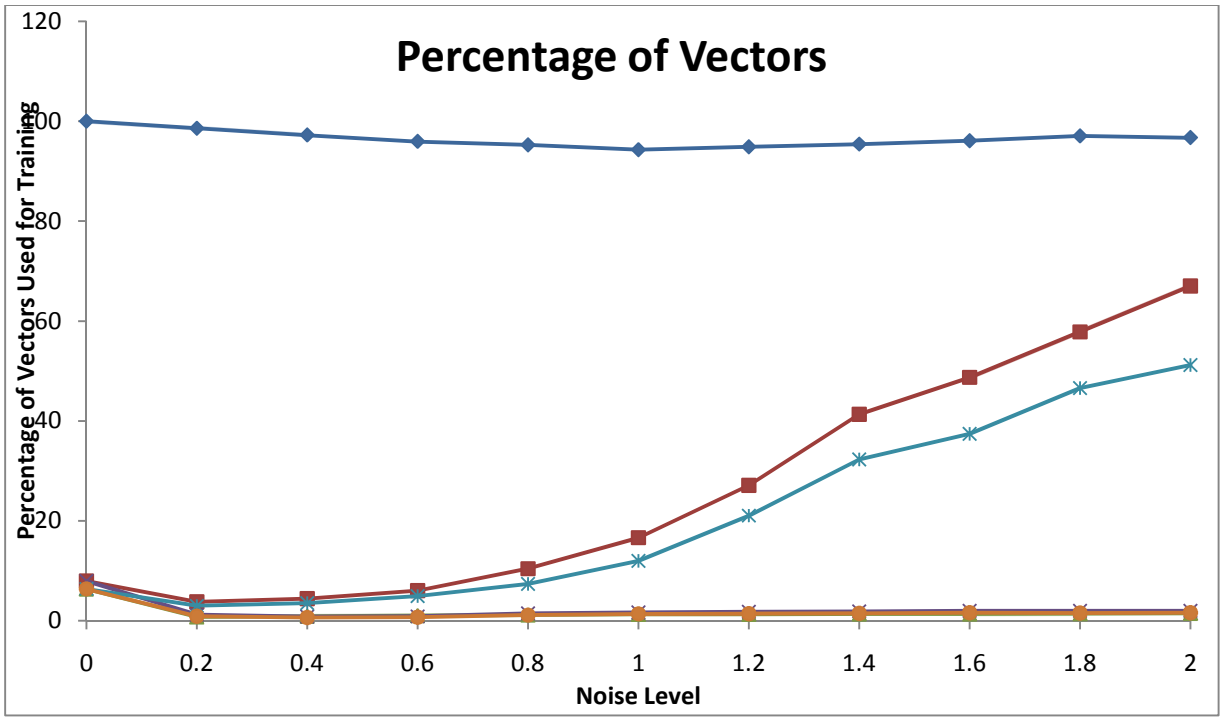

Figure 13. Percentage of vectors for simulated fMRI data

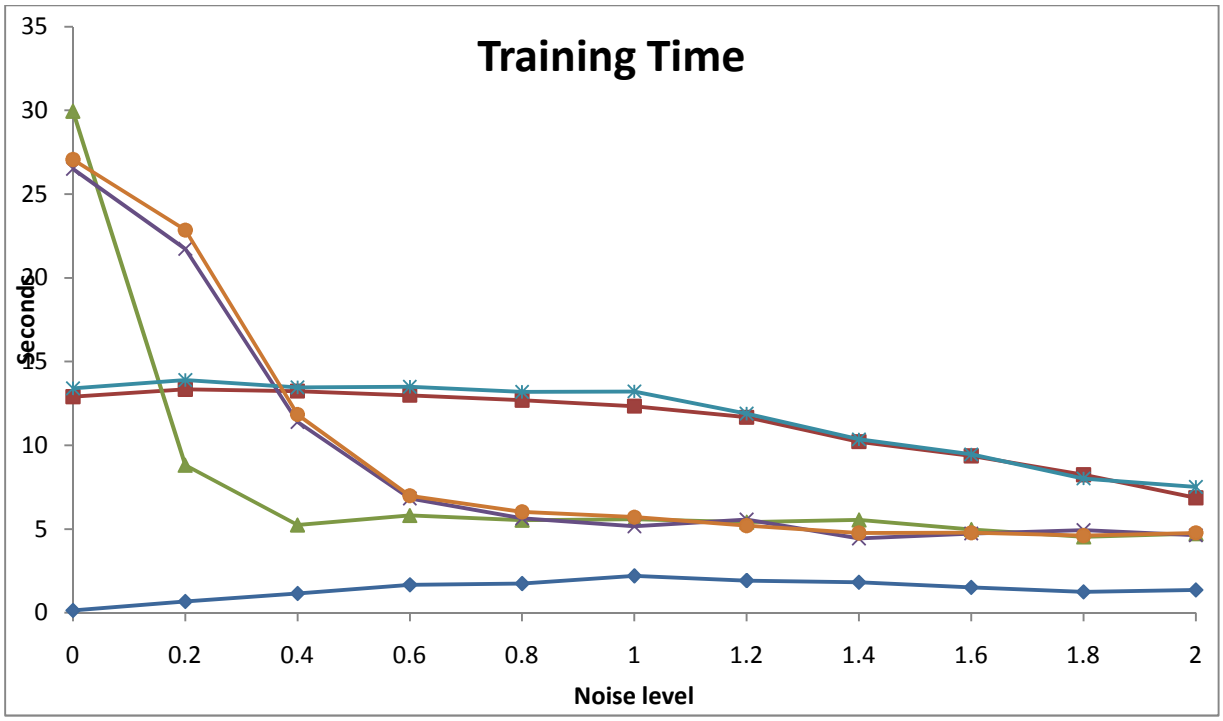

Figure 14. Training times for simulated fMRI data

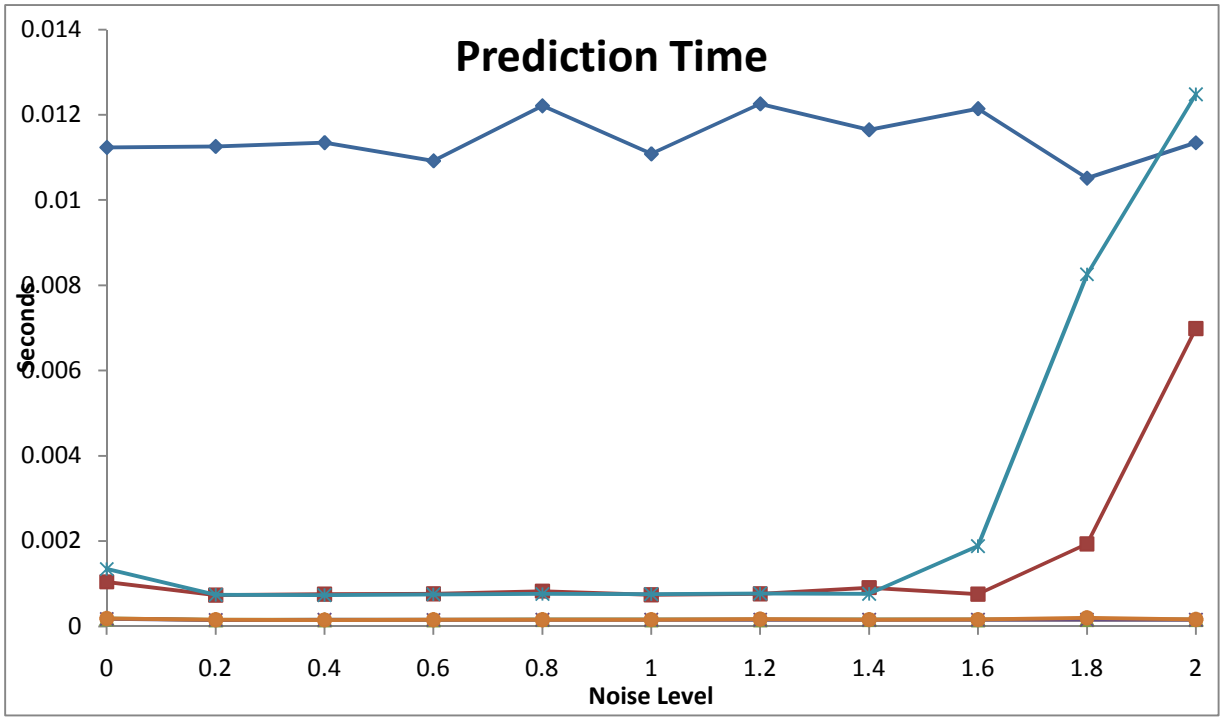

Figure 15. Prediction times for simulated fMRI data

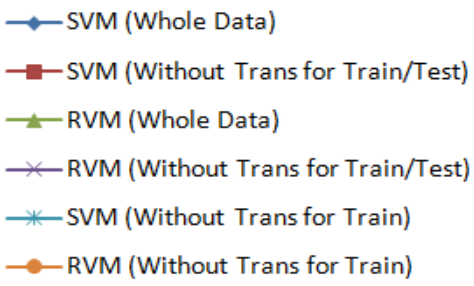

Figure 16. Legend for figures 12-15

4.2 Actual fMRI Data Results

4.2.1 Single Subject: Visual

Figure 17 reveals the performance of the two pattern recognition algorithms for a simple visual paradigm performed by a single subject. The results presented are the average of the performance metrics for 8 subjects. The accuracies are comparable between RVM and SVM. For both algorithms, accuracy increases with the use of feature selection. The removal of transitional images does not lead to an increase in accuracy. For all scenarios, RVM uses a much lower amount of vectors than SVM. Furthermore, training times and predictions times are lower for RVM than for SVM for any given condition. The removal of transitions decreases the amount of training time for both classifications. Prediction times are affected by both feature selection and the removal of transitional images. Both techniques reduce the prediction time dramatically. However, feature selection seems to have the greatest effect on this measure. For each algorithm, the best prediction speed was found to occur when both techniques were employed.

The model's weight maps are presented in Figure 18. The corresponding color bar indicates the strength of weight the two labels used. The upper half of the bar corresponds

to the visual state while the lower half corresponds with the rest state. The highest panel shows the result of the GLM analysis using a threshold of $q<0.001$. The rest of the panels show the top 10 percent of the voxels used for the model. Both classification methods correctly identified the visual cortex as most relevant to the classification task. Furthermore, the model maps show some resemblance to the GLM activation maps. It should be noted that less contrast can be observed in the RVM weight maps.

Figures 19 and 20 show the classifier outputs derived from training with all the data and training without the first three transition images in each block. It is revealed that in each case the model is trained without the transitions, it tends to lag behind the full model outputs.

Tables 1 and 2 (a-d) measure the sensitivities and precisions of the binary classification task. Sensitivity measures the proportion of actual positives which were correctly identified. In contrast, precision is a measure of accuracy provided that a specific class has been predicted. The sensitivities and precisions between RVM and SVM for any given training scenario are comparable. For both classification algorithms, feature selection is shown to improve the measures. Furthermore, the removal of transitional images yields a decrease in both measures of accuracy. For each scenario in both SVM and RVM, the classification process is more sensitive to the visual task while having a slightly higher precision rate for the resting task.

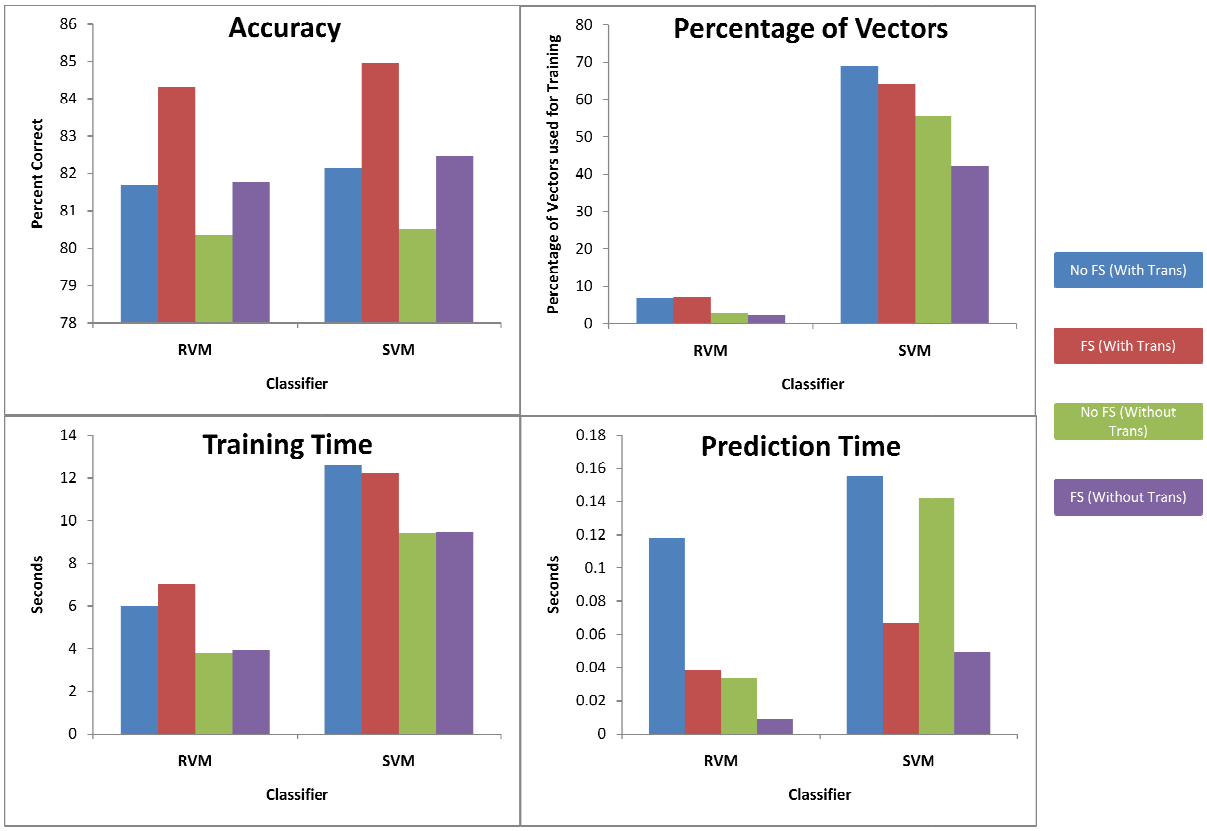

Figure 17. Single Subject: Performance measures for visual paradigm

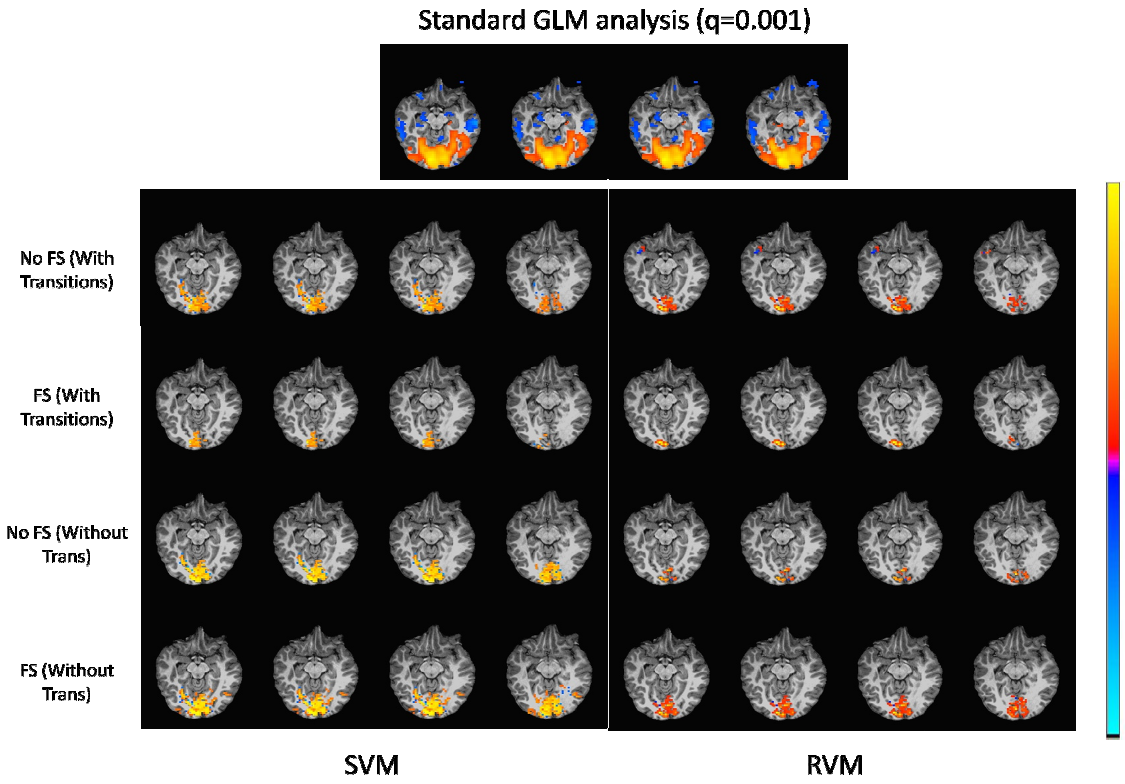

Figure 18. Weight Maps for Single Subject Visual Paradigm (Subject 3)

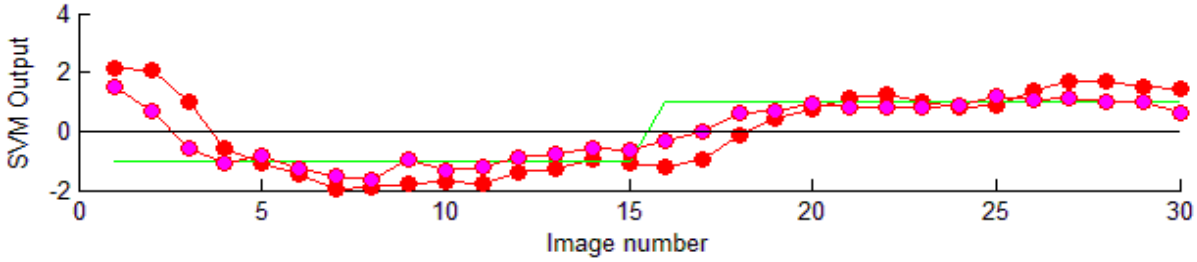

Figure 19. Sample SVM classifier output for single subject visual paradigm using feature selection. Pink denotes classification using all images while red denotes classification with removal of transitional images.

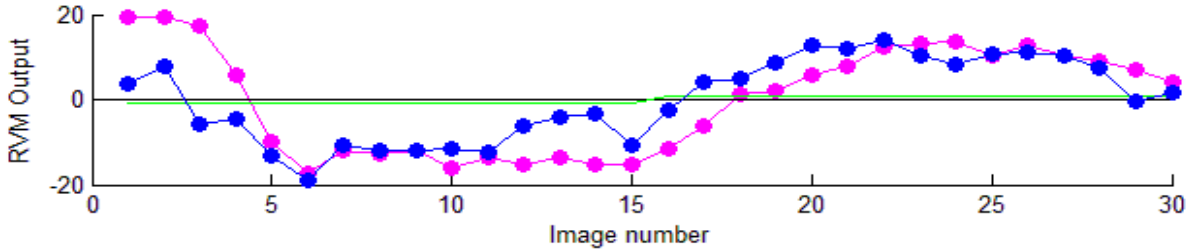

Figure 20. Sample RVM classification output for single subject visual paradigm using feature selection. Blue denotes classification using all images while magenta denotes classification with removal of transitional images.

Table 1. Confusion Matrices for RVM Classification (Subject 2)

		Whole Data/NO FS		
		Actual Value		
		Rest	Visual	Precision
Predicted Outcome	Rest	160	17	0.90
	Visual	20	163	0.89
	Sensitivity	0.89	0.91	

a.

		Whole Data/FS		
		Actual Value		
		Rest	Visual	Precision
Predicted Outcome	Rest	159	11	0.94
	Visual	21	169	0.89
	Sensitivity	0.88	0.94	

b.

		Without Transitions/No FS		
		Actual Value		
		Rest	Visual	Precision
Predicted Outcome	Rest	147	35	0.81
	Visual	33	145	0.82
	Sensitivity	0.82	0.81	

c.

		Without Transitions/FS		
		Actual Value		
		Rest	Visual	Precision
Predicted Outcome	Rest	147	28	0.84
	Visual	33	152	0.82
	Sensitivity	0.82	0.84	

d.

Table 2. Confusion Matrices for SVM Classification (Subject 2)

		Whole Data/FS		
		Actual Value		
		Rest	Visual	Precision
Predicted Outcome	Rest	154	21	0.88
	Visual	26	159	0.86
	Sensitivity	0.86	0.88	

a.

		Whole Data/FS		
		Actual Value		
		Rest	Visual	
Predicted Outcome	Rest	156	18	0.90
	Visual	24	162	0.87
	Sensitivity	0.87	0.9	

b.

		Without Transitions/No FS		
		Actual Value		
		Rest	Visual	Precision
Predicted Outcome	Rest	146	30	0.83
	Visual	34	150	0.82
	Sensitivity	0.81	0.83	

c.

		Without Transitions/FS		
		Actual Value		
		Rest	Visual	Precision
Predicted Outcome	Rest	147	26	0.85
	Visual	33	154	0.82
	Sensitivity	0.82	0.86	

d.

4.2.2 Single Subject: Motor

Figures 21 and 22 show the performance metrics for both SVM and RVM classification using right handed and left handed motor task fMRI data. The results are strikingly similar between the right and left handed motor tasks. Similar to the visual task, RVM and SVM classification accuracy was comparable. The additional preprocessing step of feature selection seems to increase accuracy for both machine learning techniques. The number of vectors used during the RVM training process was much less than the amount used in during SVM training. Training times and predictions times were lower for RVM than for SVM as well for any given scenario. In both techniques, the removal of transitional images reduces the time needed for training. The observations regarding the prediction time were identical as the one presented in the previous section. The weight maps presented in Figures 23 and 24 reveal that the classification methods correctly identified the corresponding contralateral motor cortex as relevant for the motor tasks at hand. However, the weights maps presented are not as robust as the activation maps seen using GLM analysis. Similar to the visual paradigm visualization, less contrast can be observed for the RVM case.

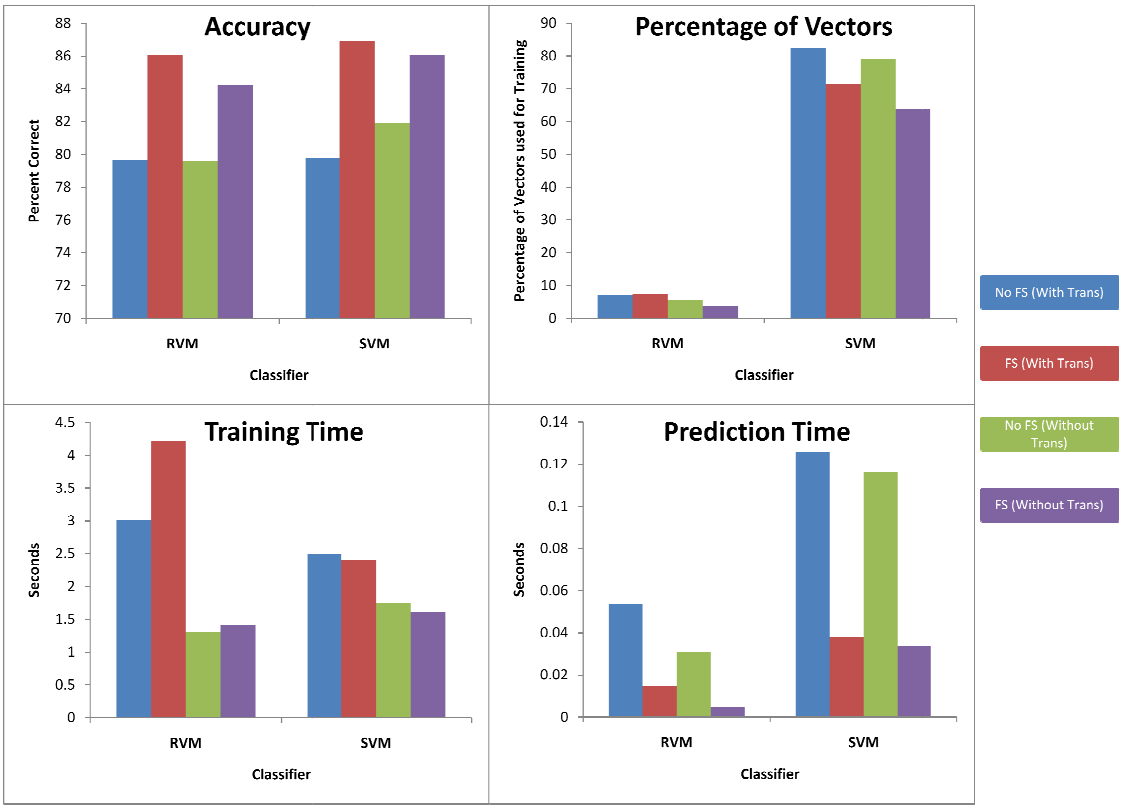

Figure 21. Single subject: Performance measures for motor right paradigm

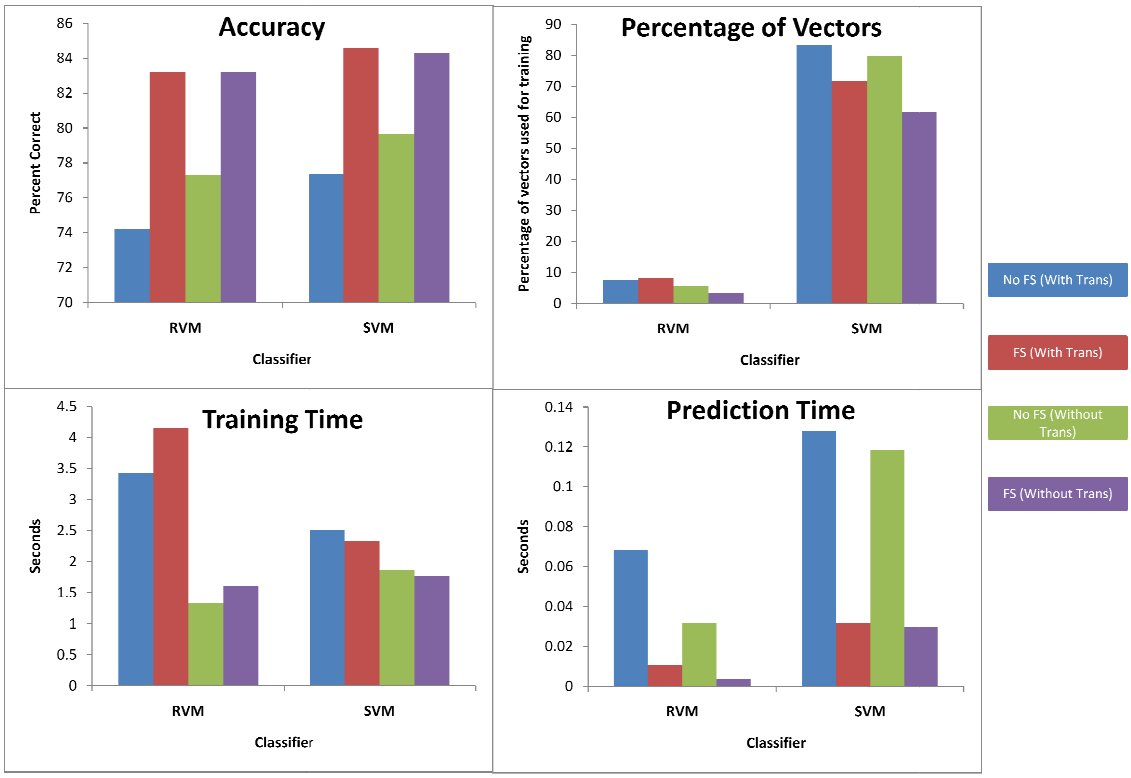

Figure 22. Single subject: Performance measures for motor left paradigm

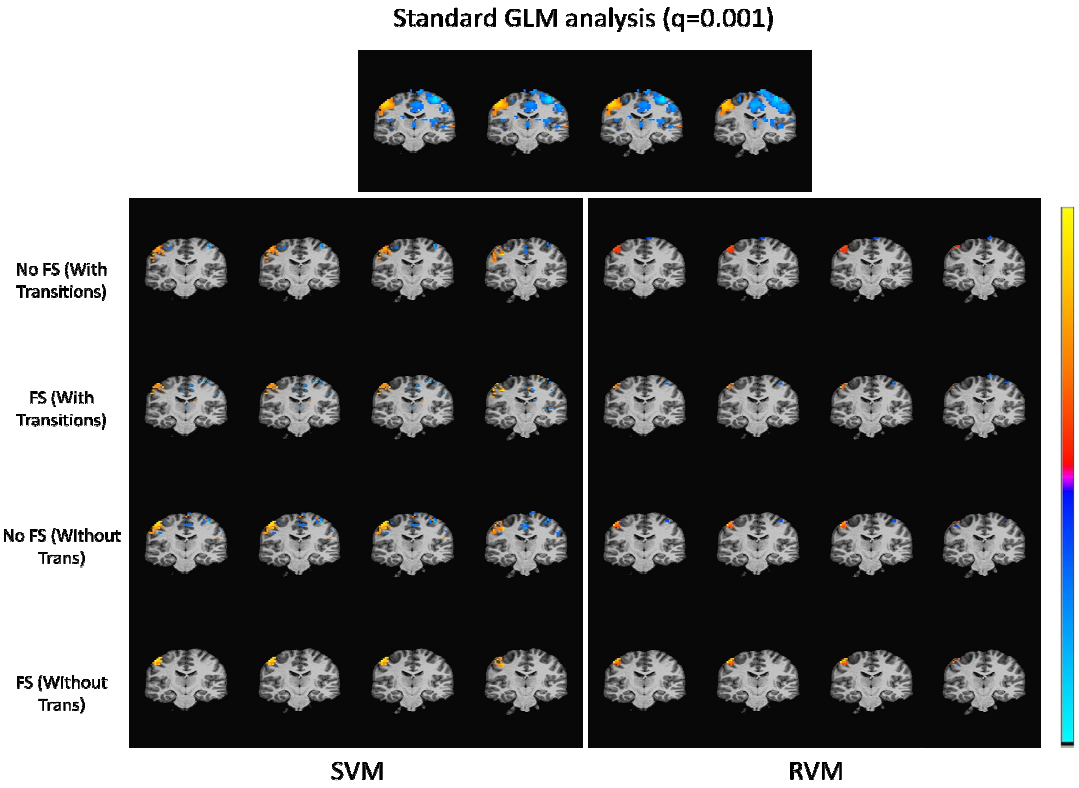

Figure 23. Weight maps for single subject motor right paradigm (Subject 4)

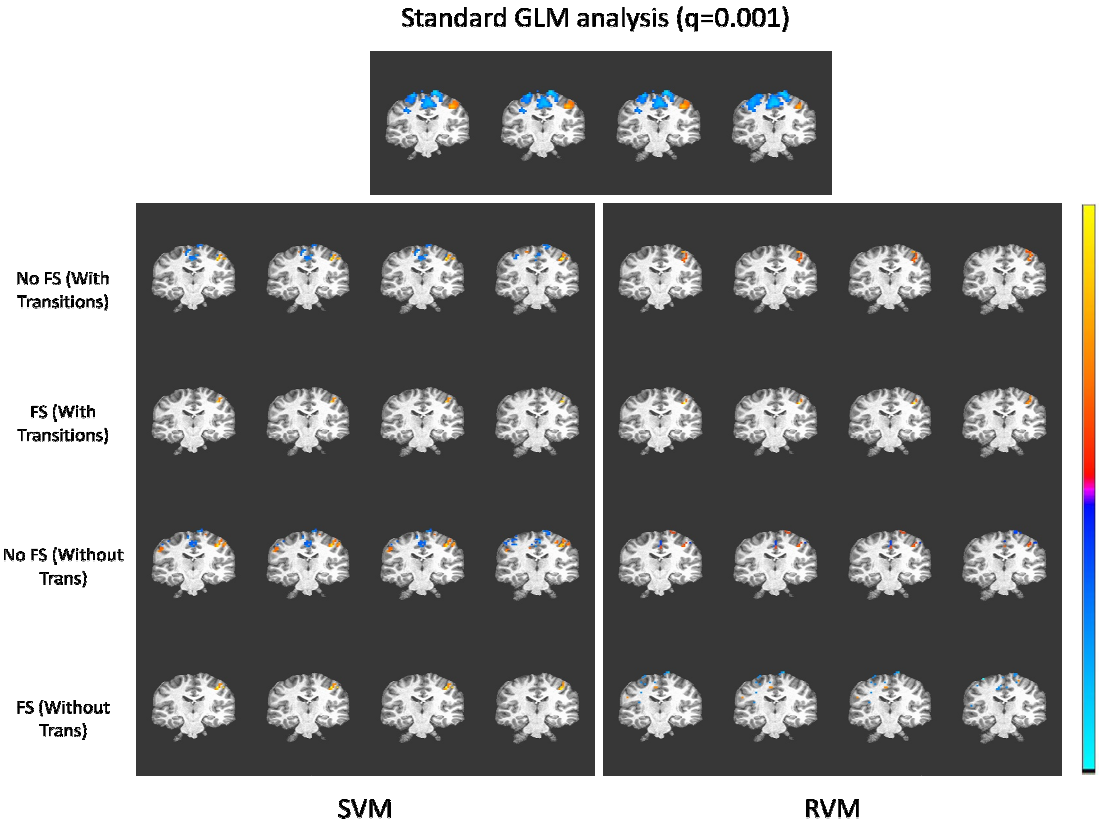

Figure 24. Weight maps for single subject motor left paradigm (Subject 4)

4.2.3 Multi-subject: Visual

In Figure 25, the performance measures of both SVM and RVM classification for a simple visual paradigm using multiple subjects are shown. In general, accuracies are lower for the multi-subject classifications than for single subject classifications. As before, accuracies between the two machine learning classifications are comparable. Feature selection tends to reduce of the number of misclassifications. The use of the RVM algorithm requires a much lower amount of vectors than the SVM algorithm. Moreover, the removal of the transitional images lowers this amount even more. For all cases, the time duration of RVM training is much less than the training time for SVM training. Again, the removal of the transitional images reduces the training time for both classification models. Multi-subject training times and prediction times are much larger than the times seen for single subject classifications. Relevance vector machine prediction time is also less than support vector machine prediction time for any given scenario. This prediction time decreases with the removal of the transitional images and with the use of feature selection.

In Figure 26, the weight maps along with a simple corresponding GLM activation map are presented. The models' weight maps for each scenario reveal that the most relevant voxels for the task is located at the visual cortex. This corresponds well with the map created with the use of the general linear model. In contrast to the single subject case, these maps are more robust and correspond more to the gold standard GLM maps. Despite this, these weight maps are a bit noisier than their single subject counterparts. As seen in the figure, this noise is reduced a bit with the removal of the transitional images for both RVM and SVM.

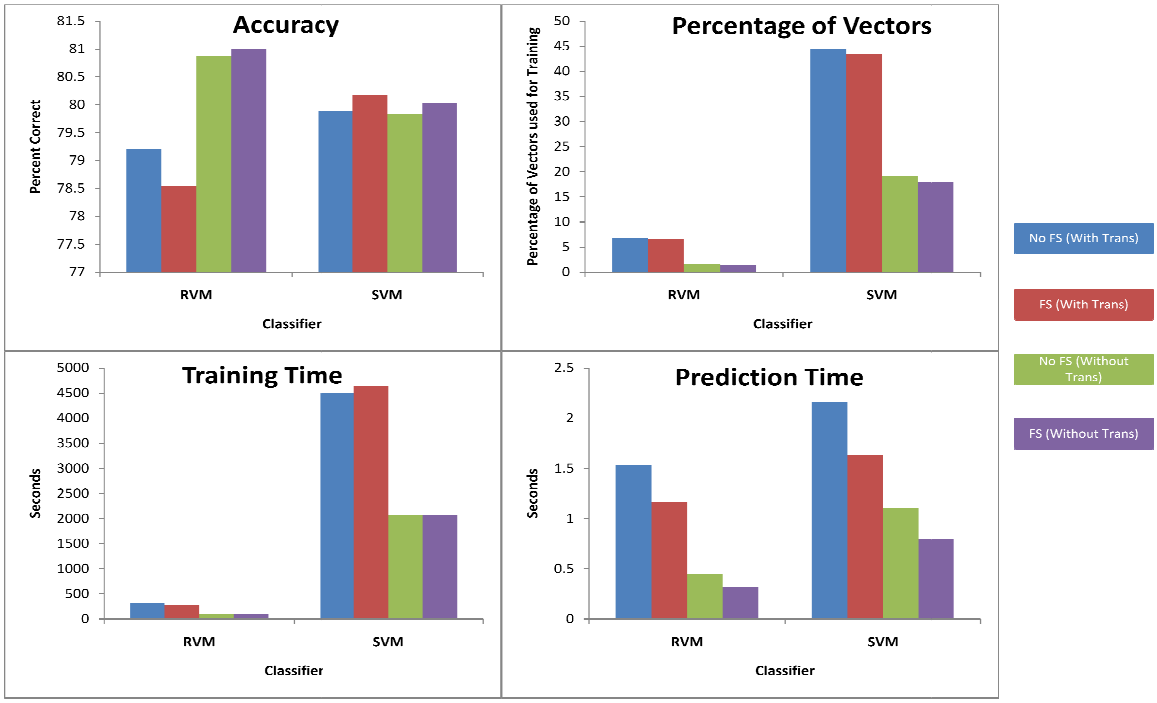

Figure 25. Multi-subject: Performance measures for visual paradigm

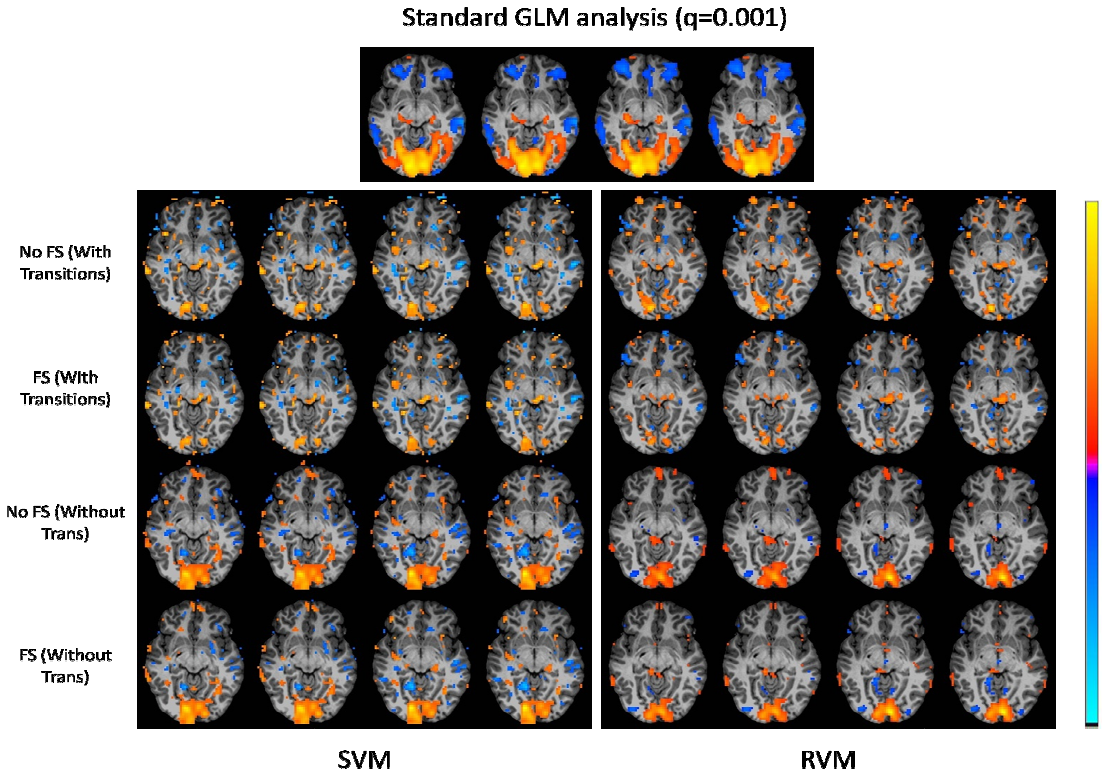

Figure 26. Weight maps for multi-subject visual paradigm

4.2.4 Multi-subject Motor

In this last section, the performance of both SVM and RVM classification on motor fMRI data was evaluated for the multi-subject case. Figures 27 and 28 show the performance measures for both classification algorithms. Accuracies tend to be much lower for the multi-subject than for single subject classifications. Similar classification accuracies were observed between the two classifications. The percentage of vectors used for training, training time, and prediction time for RVM classification were all lower than that of SVM classification. The removal of the transitional images seemed to reduce the number of vectors and training times for both algorithms. Similar to the multi-subject visual case, training times and prediction times are much than their single-subject counterparts. As seen before, the removal of transitional images coupled with the feature selection process also reduces prediction times.

Weight maps alongside with a GLM analysis map are presented in Figures 29 and 30. The top 10 percent of the weights show the most relevant voxels for the motor tasks are on the contralateral motor cortex. This corresponds well with basic neuroscience and with the GLM map shown in the uppermost panel. The removal of transitional images tends to remove the noise from the weight maps so that they resemble the GLM maps more accurately. Moreover, RVM maps reveal to have a little less contrast than the SVM maps.

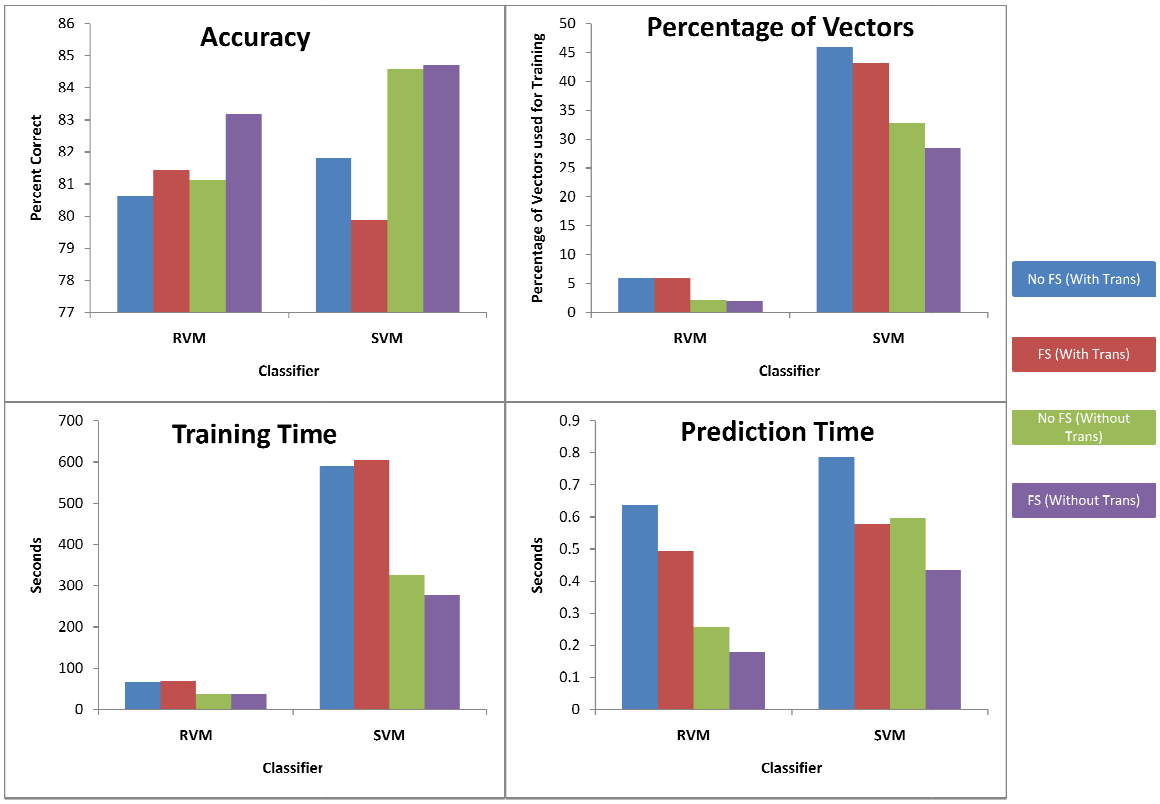

Figure 27. Multi-subject: Performance measures for motor right paradigm

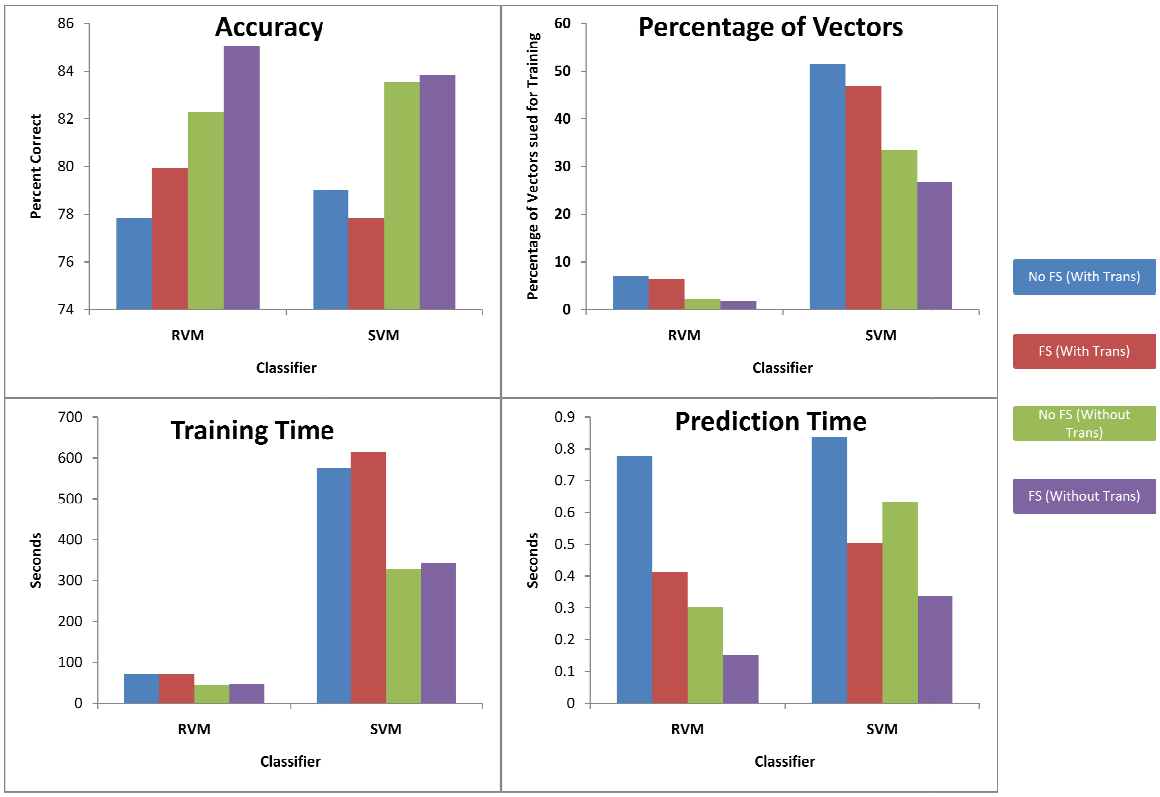

Figure 28. Multi-subject: Performance measures for motor left paradigm.

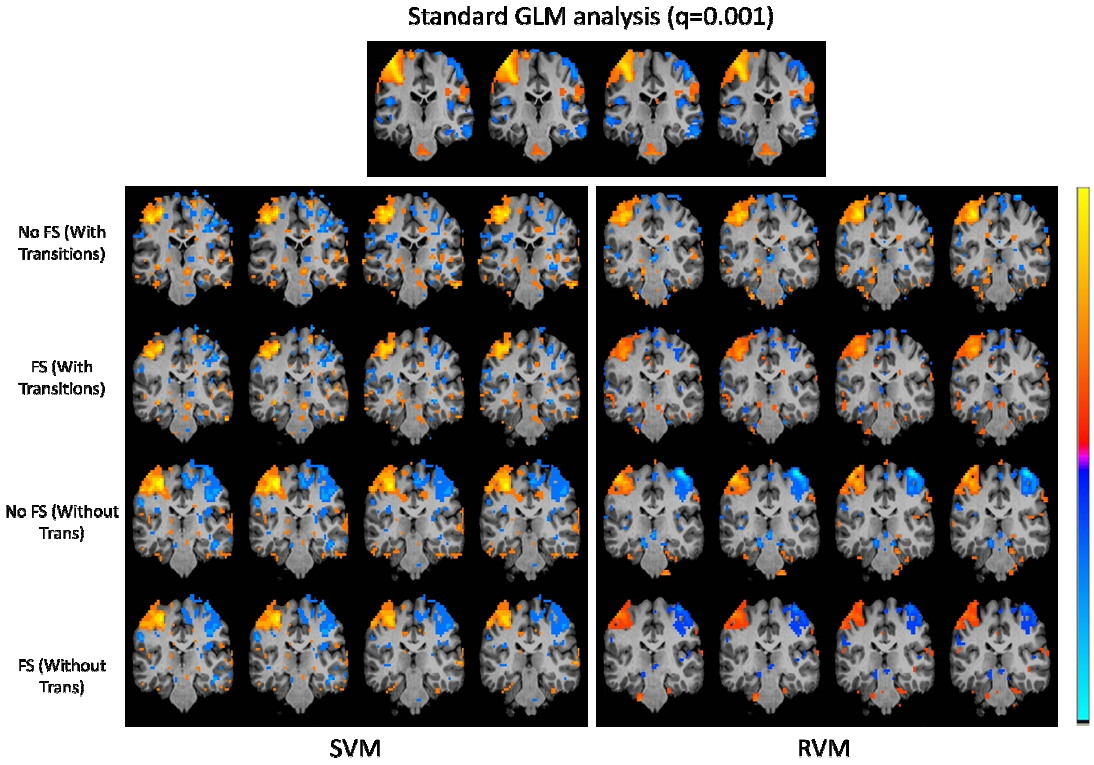

Figure 29. Weight maps for multi-subject motor right paradigm

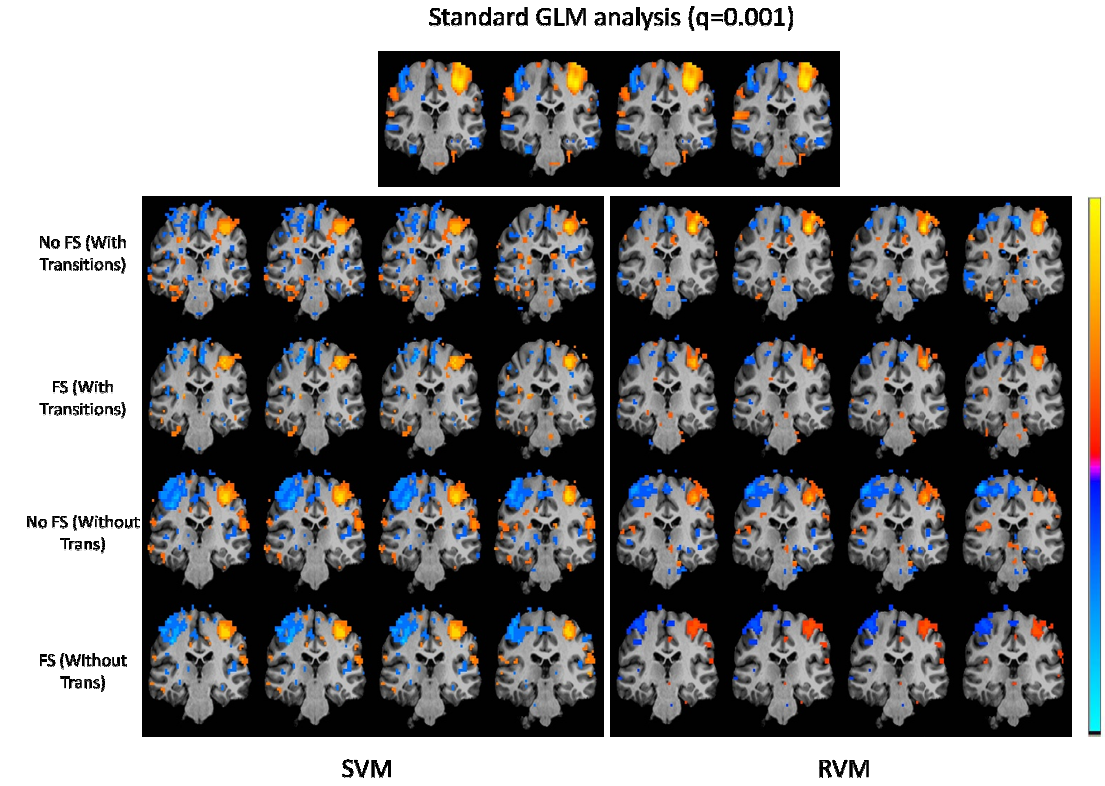

Figure 30. Weight maps for multi-subject motor left paradigm

CHAPTER 5

DISCUSSION

5.1 Simulation Results

The goal of this section was to show the effect of noise on SVM and RVM classification for simulated high dimensional fMRI data. Noise was added to a ten voxel model brain. Then, machine learning algorithms were trained and implemented in three different ways: training/testing using all of the images, training/testing with the removal of the transition images from only the training subset, and training/testing with removal of images from both sets of data. The results shown in Figures 12-15 reveal important distinctions between the two pattern recognition algorithms. Furthermore, inferences can be made for the different training/testing techniques for each case.

As seen in figure 12, the accuracy for both classification seems to decrease with the amount of noise added. When the data becomes less dissociable, there is a resulting increase in error rate. With increasing noise, the algorithms find it more difficult to distinguish between the two states. Regardless the scenario, the accuracies of both algorithms are very similar for any given noise level. The use of relevance vector machines does not sacrifice the accuracy of the prediction process.

Although the main focus of this work is to compare the performances of the two machine learning algorithms, it is interesting to note the differences between training/testing techniques. Different techniques are used depending on whether analysis is accomplished off or online. Typically for offline fMRI analysis, training and testing is

done using all of the images or by removing transitional images from both subsets of data [49]. Online studies or rtfMRI applications require training/testing using all of the images or analysis with the removal of transition images from the training set. Online studies cannot remove images from the testing subset since in practice the brain state is unknown. The presence of when the transition between two states will not be known unless the model can be trained to distinguish when this change between states occur.

In this simulation, differences in prediction accuracies can be seen in the different techniques. For any given noise level, both RVM and SVM training/testing with the removal of transitional images yields a high classification accuracy in relation to the other two techniques. With a 6 second lag resulting from the hemodynamic response, the transitional images tend to be different in character to the other two brain states. As a result, when using a two state classifier, the classifier misclassifies the transitional images. By removing these images from both subsets, the classifier truly becomes a two state predictor and can predict with high accuracy as seen in Figure 12. In the next section, differences between the two remaining techniques will be discussed in more detail.

In Figure 13, one of the main advantages of RVM classification can be seen clearly. The excitement behind using RVM over SVM stems from the use of fewer vectors during training of the model. This is clearly seen in the figure. The amount of vectors used during RVM is much less than SVM in every training/testing scenario. The increase in the number support vectors is the direct result of the increasingly nonseparability of the data. More support vectors must be implemented to train the model to create a hyperplane which effectively splits the data into two classes.

This difference in vector number could be attributed to the nature of vectors used for training in both instances. With the addition of noise, datasets become chaotic and thus increasingly inseparable. During SVM training, vectors are chosen close to the maximum margin hyperplane. RVM training involves the selection of vectors far away from the decision boundary. With inseparable data, the model for SVM will be based on data points which might have been misclassified since it's close to the decision boundary. RVM models will be inferred from datapoints far from this boundary and therefore, are more representative of the classes being studied. It should be mentioned, however, the RVM training process does not converge for extremely high noise while SVM did converge with low prediction accuracies. This nonconvergence of the RVM algorithm is commonly seen and is even brought up in the software manual written by Tipping. According to the literature, convergence depends highly on the initial kernel width used. The effect of kernel width on convergence and RVM performance merits further investigation.

At low noise, RVM training time is higher than SVM training times (Figure 14). The opposite trend could be observed as the presence of noise is increased. The discovery of low training times for both algorithms despite a corresponding high error rate is an interesting finding. This phenomenon could possibly be a result of the algorithm hitting a local minimum during the optimization process. Furthermore, for the SVM case, the parameter C set at 1 allows for many misclassifications during training. Training times might increase with a hard margin SVM using a higher C parameter. The big discrepancy in training times between low level noise and high level noise for RVM could be the result of overfitting. At no noise, overfitting seems to be present as seen by the high

number of vectors in Figurer 13. In Figure 15, the prediction times for the RVM and SVM classification is presented. In all scenarios, RVM decisions are made much faster than SVM decisions. This faster decision speed can be attributed to the RVM's sparsity.

5.2 Single Subject Results

The performance of support vector machines and relevance vector machines were compared using visual and motor fMRI data from single subjects. This comparison was done using two training techniques: training with all of the data and training with the exclusion of the transitional images. The exclusion of testing images is not included in this analysis since this part of the thesis is catered towards the possible use of these algorithms for real-time applications. In practice, the brain states during testing will be unknown and thus images depicting transitions between states cannot be extracted.

The effect of feature selection was also observed. The results in Figure 17, 21, and 22 clearly show that accuracy is not sacrificed with the use of relevance vector machines. Moreover, the amount of vectors, training time, and prediction time are all generally lower for RVM classification than for SVM classification. Lower amount of vector usage along with quicker decision times are commonly seen in literature[34,50]. The sparsity of the RVM model can yield a decision function which is much faster than its SVM counterpart. This reduction in computational complexity of the decision function makes RVM more suitable for real-time fMRI applications in which computational time is an issue.

In contrast, the observation of lower training times for RVM than for SVM is not commonly seen in literature. For the most part, RVM training takes longer than SVM for

large datasets [51]. The authors argue higher training times are seen due to complex iteration process seen during RVM model creation. In fact, the training time is proportional to the cube of the number of training samples[52]. This discrepancy in the results can be attributed to the fact that the amount of iterations for the RVM algorithm was set to its default number (same number of iterations as for the SVM algorithm).The number of iterations might have been too low and the optimization might have not complete. Regardless of this oversight, accuracies were still comparable to that of SVM.

It is apparent the removal of the transition images and the implementation of feature selection have the effect of reducing the computational complexity of the learning and prediction processes. These techniques reduce the number of datapoints. In doing so, the algorithms train with less data and output a less complex model which is used later for predictions. Despite the simplification, these techniques seem to have opposite effects in terms of prediction accuracy. Feature selection increases prediction accuracy since the model focuses only on those voxels which are known to be involved in the task at hand. This finding mirrors the results presented by Mourao-Miranda et al [47].

The removal of the transition images and its effect on the prediction accuracy for both SVM and RVM is shown in Figures 19 and 20. The accuracy seems to decrease with the removal of the transitional images. The lower accuracy is due to the increased error rate for the transition images. When the two training techniques are compared, the output from the model that is not trained with the transition images seems to be less responsive and lagging behind the full data model. The full data model seems to account for the transition effects even though they are underrepresented [53].

The confusion matrices presented in Tables 1 and 2 also show the effect of removing transitional images on measures of classification accuracy. For both classifications, the removal of the transitional images resulted in more misclassifications of both classes. Moreover, the accuracy measures (sensitivity and precision) both fell in response to this removal. This phenomenon can be attributed to the model's awareness of transitions despite its absence from the actual training labels. Feature selection improved the accuracy measures. Feature selection uses only the voxels that are pertinent to the task at hand. Therefore, since the classification focuses only the features which are implemented with the brain state, both sensitivity and precision are improved.

In all cases, sensitivity is slightly higher for the visual classification than for the resting or baseline classification. Furthermore, the opposite result is seen for the precision rate. The classification process has less false negatives for the visual prediction. The algorithm is better at recognizing the visual state. This may be due to the fact that the visual task is robust as opposed to the resting state which may involve many different brain activation patterns. However, this classification process also has more false positives for the visual task as reflected by its lower precision rate. Most misclassifications occur during the noisy transitional periods. At this juncture, the classification could either one of two ways. Since the classifiers are more sensitive to the visual task, the classifier will tend to classify these images as resulting from visual stimuli. In doing so, more false positives are observed.

The simple weight visualization scheme (Figures 18, 23, and 24) was effective in showing the voxels that were relevant for the task performed. For the visual paradigm, voxels within the visual cortex were shown to be the most relevant for the visual task. For

the motor paradigms, the contralateral motor cortex was shown to be most relevant for the motor tasks. Although some similarities can be drawn between the GLM maps and the weight maps, the weight maps contain much less voxels in the regions where activation should occur in comparison to the multi-subject weight maps (discussed in next section). This may be partly due to the fact that there are a limited amount of images used for training. Moreover, the choice of the arbitrary threshold might have eliminated voxels which may have been part of this activation. To rectify this, the threshold for the GLM maps should be thresholded to yield the same number of voxels as the weight maps. Any differences in contrast between the SVM and RVM could be the result of the location of their model vectors away from the hyperplane. Further refinements should be made to obtain weight maps that better match the patterns seen with the GLM activation maps.

5.3 Multi-Subject Results

Performance measures were also observed for the multi-subject case for both pattern recognition algorithms (Figure 25, 27, and 28). Although accuracy is generally lower for each classification in comparison to the single subject case, prediction accuracies are still greater than chance (~50%). This drop in accuracy is a direct result from the inter-subject variability typically seen in fMRI studies. As seen in the last section, accuracies are comparable between the two algorithms. The differences between the SVM and RVM performance measures were identical as the single subject analysis. Similarly, feature selection and removal of transitional images reduced training times and prediction times since there was less data used for training. The removal of the transitions actually resulted in no change or increases in prediction accuracy. This is a stark contrast

to the single subject case in which the removal of images resulted in more errors. The model becomes less 'transition' aware when a heavy volume of brain state images are used for training. It should be further noted that prediction times are much higher in the multi-subject analysis than in the single-subject analysis. An increase in training data results in a less sparse model. Therefore, the decision function is more complex and decision speeds are slower. Lower accuracies and slower predictions make multi-subject training unfavorable for real-time applications. Spatial and temporal compression techniques could be implemented to improve multi-subject classification[47].

The weight maps presented in Figures 26, 29 and 30 are much more robust than their single-subject siblings. The weight maps are more similar to the GLM activation maps due to a higher number of training samples. Noise is highly present in the maps along with the strong activation in the regions associated with the stimulus. This noise originates from the inherent intersubject variability and from the noisy transitional images which do not conform to the patterns seen during the two brain states being studied. The removal of the transitional images removes the latter source of noise and results in a cleaner weight map. Further techniques should be explored to rid the images of the noise resulting from variability between subjects.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The work presented in this thesis shows that the use of a relevance vector machine classifier is not only practical, but also a preferable alternative to support vector classification. RVM classification obtained the same prediction accuracies as SVM with a sparser model, and both a faster training and classification time. The fast computational time is of particular importance for time sensitive applications such as rtfMRI which is already limited by the hemodynamic response time. In addition to the performance measures, the weights of both algorithms were able to be visualized effectively. The voxels identified as most relevant by both methods are nearly identical, illustrating that the two methods are using the same information for classification.

Despite the possible benefits derived from using RVM for real-time fMRI studies, it still has some drawbacks. As mentioned earlier, the RVM model sometimes will not converge and can even crash. Proper parameters must be put in place depending on the application. It should be pointed out that this work involved the use of data from a simple paradigm. For most complex tasks, data separation might not be as trivial. Differences between SVM and RVM classification could vary from what was seen during this simple analysis. To my knowledge, this thesis is the first study that has attempted to create RVM weights map. In contrast, visualization of weight maps has been a highly researched topic in the field of support vector machines in fMRI analysis. Different complex schemes commonly used for SVM weight visualization could be applied to create RVM weight

maps [54]. Relevance vector regression, the use of different non-linear kernels, and multiclass classification are other possibilities worth investigating.

REFERENCES

1 Bloch, F. Nuclear Induction. *Physical Review*, 1946, 70(7-8), 460.

2 Purcell, E.M., Torrey, H.C. and Pound, R.V. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Physical Review*, 1946, 69(1-2), 37.

3 Damadian, R. Tumor Detection by Nuclear Magnetic Resonance. *Science*, 1971, 171(3976), 1151-1153.

4 Lauterbur, P.C. Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance. *Nature*, 1973, 242(5394), 190-191.

5 Kumar, A., Welti, D. and Ernst, R.R. NMR Fourier zeugmatography. *Journal of Magnetic Resonance (1969)*, 1975, 18(1), 69-83.

6 Mansfield, P. Multi-planar image formation using NMR spin echoes. *Journal of Physics C: Solid State Physics*, 1977(3), L55.

7 Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, B.P., Kennedy, D.N., Hoppel, B.E., Cohen, M.S. and Turner, R. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences of the United States of America*, 1992, 89(12), 5675-5679.

8 Peter, A.B., Eric, C.W., Hinks, R.S., Ronald, S.T. and James, S.H. Time course EPI of human brain function during task activation. *Magnetic Resonance in Medicine*, 1992, 25(2), 390-397.

9 Donald W. McRobbie, E.A.M., Martin J. Graves, Martin R. Prince. *MRI: From Picture to Proton*. (Cambridge University Press, Cambridge, U.K., 2006).

10 E. Mark Haacke, R.W.B., Michael R. Thompson, Ramesh Venkatesan. *Magnetic Resonance Imaging: Physical principles and Sequence Design*. (John Wiley and Sons, Inc. , New York, 1999).

11 Ogawa, S., Lee, T.M., Kay, A.R. and Tank, D.W. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 1990, 87(24), 9868-9872.

12 Logothetis, N.K. What we can do and what we cannot do with fMRI. *Nature*, 2008, 453(7197), 869-878.

13 Poellinger, A., Thomas, R., Lio, P., Lee, A., Makris, N., Rosen, B.R. and Kwong, K.K. Activation and Habituation in Olfaction--An fMRI Study. *NeuroImage*, 2001, 13(4), 547-560.

14 James, G.A.a., Li, X.c., DuBois, G.E.b., Zhou, L.a. and Hu, X.P.a. Prolonged insula activation during perception of aftertaste. *Neuroreport*, 2009, 20(3), 245-250.

15 Amaro, J.E. and Barker, G.J. Study design in fMRI: Basic principles. *Brain and Cognition*, 2006, 60(3), 220-232.

16 Constable, Todd, R., Spencer and Dennis, D. *Repetition time in echo planar functional MRI*. (Williams & Wilkins, Baltimore, MD, ETATS-UNIS, 2001).

17 Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D. and Frackowiak, R.S.J. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 1994, 2(4), 189-210.

18 Rowe, Daniel, B., Hoffmann and Raymond, G. *Multivariate statistical analysis in fMRI*. (Institute of Electrical and Electronics Engineers, New York, NY, ETATS-UNIS, 2006).

19 Cox, D.D. and Savoy, R.L. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 2003, 19(2), 261-270.

20 Robert, W.C., Andrzej, J. and James, S.H. Real-Time Functional Magnetic Resonance Imaging. *Magnetic Resonance in Medicine*, 1995, 33(2), 230-236.

21 Bakker, L.N. *Brain Mapping Research Developments*. (Nova Science Publishers, New York, 2008).

22 deCharms, R.C., Maeda, F., Glover, G.H., Ludlow, D., Pauly, J.M., Soneji, D., Gabrieli, J.D.E. and Mackey, S.C. Control over brain activation and pain learned by using real-time functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102(51), 18626-18631.

23 Lebedev, M.A. and Nicolelis, M.A.L. Brain-machine interfaces: past, present and future. *Trends in Neurosciences*, 2006, 29(9), 536-546.

24 Christopher deCharms, R. Applications of real-time fMRI. *Nat Rev Neurosci*, 2008, 9(9), 720-729.

25 Pereira, F., Mitchell, T. and Botvinick, M. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 2009, 45(1, Supplement 1), S199-S209.

26 Haynes, J.-D. and Rees, G. Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, 2006, 7(7), 523-534.

27 Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 1995, 20(3), 273-297.

28 Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 1998, 2(2), 121-167.

29 Ben-Hur, A., Ong, C.S., Sonnenburg, S.r., Schalkopf, B. and RÃ¤tsch, G. Support Vector Machines and Kernels for Computational Biology. *PLoS Comput Biol*, 2008, 4(10), e1000173.

30 Bernhard, E.B., Isabelle, M.G. and Vladimir, N.V. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory* (ACM, Pittsburgh, Pennsylvania, United States, 1992).

31 Liu, J., Ranka, S. and Kahveci, T. Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*, 2008, 24(13), i86-95.

32 Ganapathiraju, Aravind, Hamaker, Jonathan, E., Picone and Joseph. *Applications of support vector machines to speech recognition*. (Institute of Electrical and Electronics Engineers, New York, NY, ETATS-UNIS, 2004).

33 Craddock, R.C., Holtzheimer, P.E., Hu, X.P. and Mayberg, H.S. Disease State Prediction from Resting State FMRI. *NeuroImage*, 2009, 47(Supplement 1), S80-S80.

34 Tipping, M.E. The Relevance Vector Machine. *Advances in Neural Information Processing Systems*, 2000, 12, 652-658.

35 Tipping, M.E. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 2001, 1, 411-444.

36 MacKay, D.J.C. Bayesian Interpolation. *Neural Computation*, 1992, 4(3), 415-447.

37 Nabney, I.T. Efficient training of RBF networks for classification. *IEE Conference Publications*, 1999, 1999(CP470), 210-215.

38 Liyang, W., Yongyi, Y., Nishikawa, R.M., Wernick, M.N. and Edwards, A. Relevance vector machine for automatic detection of clustered microcalcifications. *Medical Imaging, IEEE Transactions on*, 2005, 24(10), 1278-1285.

39 Li, Y., Campbell, C. and Tipping, M. Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, 2002, 18(10), 1332-1339.

40 Flake, J., Moon, T.K., McKee, M. and Gunther, J.H. Application of the relevance vector machine to canal flow prediction in the Sevier River Basin. *Agricultural Water Management*, 97(2), 208-214.

41 Maurice Hollman, S.B., Charles Mueller, Johannes Bernarding. Predicting Human Decisions in a Social Interaction-Scenario Using Real-Time Functional Magnetic Resonance Imaging (Rt-FMRI). *International Society for Magnetic Resonance in Medicine* Hawaii, 2009).

42 Robert, W.C. and James, S.H. Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, 1997, 10(4-5), 171-178.

43 LaConte, S., Strother, S., Cherkassky, V., Anderson, J. and Hu, X. Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, 2005, 26(2), 317-329.

44 LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D. and Strother, S. The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics. *NeuroImage*, 2003, 18(1), 10-27.

45 Logothetis, N.K. and Wandell, B.A. INTERPRETING THE BOLD SIGNAL. *Annual Review of Physiology*, 2004, 66(1), 735-774.

46 Jean Talairach, P.T. *Co-Planar Stereotaix Atlas of the Human Brain: 3-D Proportional System: An Approach to Cerebral Imaging*. (Thieme Medical Publishers, Inc. , New York, 1988).

47 Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G. and Brammer, M. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage*, 2006, 33(4), 1055-1065.

48 Hardoon, D., Mourão-Miranda, J., Brammer, M. and Shawe-Taylor, J. Using Image Stimuli to Drive fMRI Analysis. *Neural Information Processing*, pp. 477-4862008).

49 Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M. and Newman, S. Learning to Decode Cognitive States from Brain Images. *Machine Learning*, 2004, 57(1), 145-175.

50 Dimitris G. Tzikas, L.W., Aristidis Likas, Yongyi Yang, P Galatsanos. A Tutorial on Relevance Vector Machines for Regression and Classification with Applications. University of Ioannina, Greece, 2006).

51 Oliver, W. Sparse Bayesian Learning for Efficient Visual Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27, 1292-1304.

52 SF Wong, R.C. Real-time Interpretation of Hand Motions using a Sparse Bayesian Classifier on Motion Gradient Orientation Images. *Proceedings of the British Machine Vision*2005).

53 Stephen, M.L., Scott, J.P. and Xiaoping, P.H. Real-time fMRI using brain-state classification. *Human Brain Mapping*, 2007, 28(10), 1033-1044.

54 Mourão-Miranda, J., Friston, K.J. and Brammer, M. Dynamic discrimination analysis: A spatial-temporal SVM. *NeuroImage*, 2007, 36(1), 88-99.